Neural Networks and Intellect: Using Model-Based Concepts

Leonid I. Perlovsky

OXFORD UNIVERSITY PRESS

VISIT…

Neural Networks and Intellect

This page intentionally left blank

Neural Networks and Intellect

Using Model-Based Concepts

Leonid I. Perlovsky

New York • Oxford
OXFORD UNIVERSITY PRESS
2001

Oxford University Press

Oxford New York
Athens Auckland Bangkok Bogotá Buenos Aires Calcutta
Cape Town Chennai Dar es Salaam Delhi Florence Hong Kong Istanbul
Karachi Kuala Lumpur Madrid Melbourne Mexico City Mumbai
Nairobi Paris São Paulo Shanghai Singapore Taipei Tokyo Toronto Warsaw

and associated companies in
Berlin Ibadan

Copyright © 2001 by Oxford University Press, Inc.

Published by Oxford University Press, Inc.,
198 Madison Avenue, New York, New York, 10016
http://www.oup-usa.org

Oxford is a registered trademark of Oxford University Press

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
electronic, mechanical, photocopying, recording, or otherwise,
without the prior permission of Oxford University Press.

Library of Congress Cataloging-in-Publication Data

Perlovsky, Leonid I.
Neural networks and intellect : using model-based concepts / Leonid I. Perlovsky.
p. cm.
Includes bibliographical references and index.
ISBN 0-19-511162-1
1. Neural networks (computer science) 2. Mathematical models. I. Title.
QA76.87.P435 2000
006.3'2—dc21 00-026297

Printing (last digit): 9 8 7 6 5 4 3 2 1

Printed in the United States of America
on acid-free paper

CONDENSED TABLE OF CONTENTS

PREFACE xix

PART ONE: OVERVIEW: 2300 YEARS OF PHILOSOPHY, 100 YEARS OF MATHEMATICAL LOGIC, AND 50 YEARS OF COMPUTATIONAL INTELLIGENCE

1 Introduction: Concepts of Intelligence 3

2 Mathematical Concepts of Mind 51

3 Mathematical versus Metaphysical Concepts of Mind 125

PART TWO: MODELING FIELD THEORY: NEW MATHEMATICAL THEORY OF INTELLIGENCE WITH ENGINEERING APPLICATIONS

4 Modeling Field Theory 153

5 MLANS: Maximum Likelihood Adaptive Neural System for Grouping and Recognition 206

6 Einsteinian Neural Network 263

7 Prediction, Tracking, and Dynamic Models 289

8 Quantum Modeling Field Theory (QMFT) 321

9 Fundamental Limitations on Learning 329

10 Intelligent System Organization: MFT, Genetic Algorithms, and Kant 356

PART THREE: FUTURISTIC DIRECTIONS: FUN STUFF: MIND—PHYSICS + MATHEMATICS + CONJECTURES

11 Gödel Theorems, Mind, and Machine 383

12 Toward Physics of Consciousness 391

LIST OF SYMBOLS 425

DEFINITIONS 429

BIBLIOGRAPHY 447

INDEX 461

This page intentionally left blank

CONTENTS

PREFACE xix

PART ONE: OVERVIEW: 2300 YEARS OF PHILOSOPHY, 100 YEARS OF MATHEMATICAL LOGIC, AND 50 YEARS OF COMPUTATIONAL INTELLIGENCE

1 Introduction: Concepts of Intelligence 3

1.1 CONCEPTS OF INTELLIGENCE IN MATHEMATICS, PSYCHOLOGY, AND PHILOSOPHY **3**

1.1.1 What Is Intelligence? **3**

1.1.2 Plato, Occam, and Neural Networks **4**

1.1.3 Rule-Based Artificial Intelligence, Complexity, and Aristotle **6**

1.1.4 Philosophy vs. Architecture of Intelligent Tracker **8**

1.1.5 Summary **12**

1.2 PROBABILITY, HYPOTHESIS CHOICE, PATTERN RECOGNITION, AND COMPLEXITY **13**

1.2.1 Prerequisite: Basic Notions of the Theory of Probability **13**

1.2.2 Classical Hypotheses Choice Paradigms and Definitions **20**

1.2.3 Pattern Recognition **22**

1.2.4 A Priori Information and Adaptation **24**

1.2.5 Mathematical Formulation of Model-Based Recognition **27**

1.2.6 Conundrum of Combinatorial Complexity **29**

1.3 PREDICTION, TRACKING, AND DYNAMIC MODELS **29**

1.3.1 Linear Regression **30**

1.3.2 Regression as an Expectation **32**

1.3.3 Autoregression **33**

1.3.4 Tracking **35**

1.3.5 Association Problem **37**

1.4 PREVIEW: INTELLIGENCE, INTERNAL MODEL, SYMBOL, EMOTIONS, AND CONSCIOUSNESS **42**
Notes **45**
Bibliographical Notes **46**
Problems **47**

2 Mathematical Concepts of Mind 51

2.1 COMPLEXITY, ARISTOTLE, AND FUZZY LOGIC **52**
2.1.1 Conundrum of Combinatorial Complexity **52**
2.1.2 Adaptivity, Apriority, and Complexity **53**
2.1.3 Fuzzy Logic and Complexity **55**

2.2 NEAREST NEIGHBORS AND DEGENERATE GEOMETRIES **58**
2.2.1 The Nearest Neighbor Concept **58**
2.2.2 Mathematical Formulation **59**
2.2.3 What Constitutes Simple and Complex Classification Problems? **59**
2.2.4 Degenerate Geometry of Classification Spaces **60**

2.3 GRADIENT LEARNING, BACK PROPAGATION, AND FEEDFORWARD NEURAL NETWORKS **62**
2.3.1 Concept of Discriminating Surfaces and Gradient Learning **62**
2.3.2 Mathematical Formulation **64**
2.3.3 Learning Disability **67**

2.4 RULE-BASED ARTIFICIAL INTELLIGENCE **68**
2.4.1 Minsky, Apriority, and Adaptivity **68**
2.4.2 Soar Production System **70**

2.5 CONCEPT OF INTERNAL MODEL **73**
2.5.1 Prolegomena: Parametric vs. Nonparametric Estimation **73**
2.5.2 Model-Based Vision (MBV) **74**
2.5.3 Adaptivity and MBV **75**

2.6 ABDUCTIVE REASONING **76**
2.6.1 Deduction, Induction, and Abduction **76**
2.6.2 Abductive Reasoning Trees and Bayesian Networks **77**

2.7 STATISTICAL LEARNING THEORY AND SUPPORT VECTOR MACHINES **79**
2.7.1 Model Complexity: Risk Minimization vs. PDF Estimation **79**

2.7.2 Consistency of ERM and VC Dimension **81**
2.7.3 Support Vector Machines (SVM) **82**
2.8 AI DEBATES PAST AND FUTURE **85**
2.8.1 Arguments and Disagreements: An Overview **85**
2.8.2 Can a Machine Think? **87**
2.8.3 Rule-Based AI vs. Connectivism **90**
2.8.4 Emerging Debates **91**
2.9 SOCIETY OF MIND **94**
2.9.1 Society of Agents **94**
2.9.2 Types of Agents **95**
2.9.3 Frames and Unity of Apperception **96**
2.9.4 Limitations and What Is Next **96**
2.10 SENSOR FUSION AND JDL MODEL **97**
2.10.1 Sensor Fusion and Origins of JDL Model **97**
2.10.2 Definitions, Issues, and Types of Fusion Problems **98**
2.10.3 Sensor Fusion Levels **99**
2.10.4 Hierarchy of JDL Model Organization **100**
2.11 HIERARCHICAL ORGANIZATION **100**
2.12 SEMIOTICS **104**
2.13 EVOLUTIONARY COMPUTATION, GENETIC ALGORITHMS, AND CAS **106**
2.13.1 Complex Adaptive Systems (CAS) **107**
2.13.2 CAS: Complexity vs. Fuzziness **109**
2.14 NEURAL FIELD THEORIES **110**
2.14.1 Grossberg's Method: Physics of Mind **110**
2.14.2 ART Neural Network **111**
2.14.3 Illusions and A Priori Contents of Vision **114**
2.14.4 Motor Coordination and Sensorimotor Control **115**
2.14.5 Emotions and Learning **116**
2.14.6 Quantum Neurodynamics **118**
2.14.7 Modeling Field Theory **119**
2.15 INTELLIGENCE, LEARNING, AND COMPUTABILITY **120**
2.15.1 Computability: Turing vs. Physics **120**
2.15.2 Computational Methods of Intelligence: Summary **121**
Notes **121**
Bibliographical Notes **122**
Problems **124**

3 Mathematical versus Metaphysical Concepts of Mind 125

3.1 PROLEGOMENON: PLATO, ANTISTHENES, AND ARTIFICIAL INTELLIGENCE **126**

3.2 LEARNING FROM ARISTOTLE TO MAIMONIDES **127**
- 3.2.1 The Controversy of Aristotle **127**
- 3.2.2 Finite Angels of Maimonides **128**
- 3.2.3 Nexus of Aquinas **131**

3.3 HERESY OF OCCAM AND SCIENTIFIC METHOD **131**
- 3.3.1 Cynics, Occam, and Empiricism **131**
- 3.3.2 Nominalism, Behaviorism, and Cybernetics **132**

3.4 MATHEMATICS VS. PHYSICS **135**
- 3.4.1 Pythagoras, Descartes, Newton **135**
- 3.4.2 Computation: Metaphor vs. Physical Model **137**
- 3.4.3 Physics of Mind vs. Physics of Brain **138**

3.5 KANT: PURE SPIRIT AND PSYCHOLOGY **138**

3.6 FREUD VS. JUNG: PSYCHOLOGY OF PHILOSOPHY **140**

3.7 WITHER WE GO FROM HERE? **141**
- 3.7.1 Apriority and Adaptivity **141**
- 3.7.2 Fuzzy Logic, Models, and Neural Fields **143**

Notes **145**

Bibliographical Notes **147**

Problems **148**

PART TWO: MODELING FIELD THEORY: NEW MATHEMATICAL THEORY OF INTELLIGENCE WITH ENGINEERING APPLICATIONS

4 Modeling Field Theory 153

4.1 INTERNAL MODELS, UNCERTAINTIES, AND SIMILARITIES **154**
- 4.1.1 Certainty and Uncertainties **154**
- 4.1.2 Models and Levels **154**
- 4.1.3 Lower Level Models **155**
- 4.1.4 Similarity Measures **156**

4.2 MODELING FIELD THEORY DYNAMICS **160**
- 4.2.1 Overview of the MFT System **161**
- 4.2.2 MFT Dynamic Equations **162**
- 4.2.3 Continuous MFT System **163**

4.2.4 Heterarchy, Multiple Scales, and Local Maxima **163**
4.2.5 MFT, Fuzzy Logic, and Aristotelian Forms **164**

4.3 BAYESIAN MFT **165**
4.3.1 Bayesian A-Similarity Measure and the Principle of Maximum Likelihood **165**
4.3.2 Bayesian AZ-Similarity Measure **168**
4.3.3 MLANS Learning Equations **169**

4.4 SHANNON–EINSTEINIAN MFT **172**
4.4.1 Einstein, Likelihood, and Electromagnetic Spectrum **172**
4.4.2 Einsteinian Gaussian Mixture Model **174**
4.4.3 Equilibrium of the Photon Ensemble **176**
4.4.4 Einsteinian Likelihood and Shannon's Mutual Information **177**
4.4.5 Information and Alternative Choice States **177**
4.4.6 Mutual Model-Data Information **179**
4.4.7 Shannon–Einsteinian Similarity **181**
4.4.8 Shannon–Einsteinian MFT Dynamics **183**
4.4.9 Historical Roots of Maximum Information and Maximum Entropy Estimation **185**
4.4.10 Likelihood, Information, Ergodicity, and Uncertainty **186**
4.4.11 Forward and Inverse Problems **186**

4.5 MODELING FIELD THEORY NEURAL ARCHITECTURE **187**

4.6 CONVERGENCE **189**
4.6.1 Aspects of Convergence **189**
4.6.2 Proof of Convergence **190**

4.7 LEARNING OF STRUCTURES, AIC, AND SLT **192**

4.8 INSTINCT OF WORLD MODELING: KNOWLEDGE INSTINCT **194**

4.9 SUMMARY **194**
Notes **196**
Bibliographical Notes **198**
Problems **198**

5 MLANS: Maximum Likelihood Adaptive Neural System for Grouping and Recognition 206

5.1 GROUPING, CLASSIFICATION, AND MODELS **206**

5.2 GAUSSIAN MIXTURE MODEL: UNSUPERVISED LEARNING OR GROUPING **208**
5.2.1 Architecture and Parameters **208**

5.2.2 Likelihood Structure and Learning Algorithm **210**
5.2.3 Examples of MLANS Unsupervised Classification **213**
5.3 COMBINED SUPERVISED AND UNSUPERVISED LEARNING **225**
5.3.1 Supervised and Unsupervised Learning **225**
5.3.2 Perfect Teacher **227**
5.3.3 Probabilistic or Fuzzy Teacher **227**
5.3.4 Partial Supervision **228**
5.3.5 Examples **229**
5.4 STRUCTURE ESTIMATION **231**
5.4.1 Goals and Approaches of Structural Optimization: Models vs. Decisions **231**
5.4.2 Maximum Likelihood Estimation of Structure **233**
5.4.3 Minimum Classification Entropy **234**
5.4.4 Other Structural Issues **236**
5.5 WISHART AND RICIAN MIXTURE MODELS FOR RADAR IMAGE CLASSIFICATION **238**
5.5.1 Synthetic Aperture Radar **238**
5.5.2 Data Description **239**
5.5.3 Physically Based Clutter and Target Models **241**
5.5.4 NASA Data Examples **245**
5.5.5 Stockbridge Data Examples **246**
5.5.6 Summary of SAR Models **250**
5.6 CONVERGENCE **250**
5.6.1 Convergence and Learning **250**
5.6.2 The ML Equations **252**
5.6.3 Local Convergence and EM Algorithm **254**
5.6.4 Global Convergence **256**
5.7 MLANS, PHYSICS, BIOLOGY, AND OTHER NEURAL NETWORKS **257**
Note **260**
Bibliographical Notes **260**
Problems **261**

6 Einsteinian Neural Network 263

6.1 IMAGES, SIGNALS, AND SPECTRA **263**
6.1.1 Definitions, Notations, and Simple Signal Models **263**
6.1.2 Frequency Components, Spectrum, and Spectral Models **265**
6.1.3 Model-Based Spectrum Estimation **269**

6.2 SPECTRAL MODELS **269**

6.3 NEURAL DYNAMICS OF ENN **271**

6.3.1 Shannon's Similarity Dynamics of Einsteinian Spectral Models **271**

6.3.2 Two-Dimensional Time–Frequency ENN **272**

6.4 APPLICATIONS TO ACOUSTIC TRANSIENT SIGNALS AND SPEECH RECOGNITION **274**

6.4.1 Transient Signals **274**

6.4.2 Examples of One-Dimensional Spectrum Estimation **274**

6.4.3 Two-Dimensional Time–Frequency Models **278**

6.4.4 Hierarchical ENN + MLANS Architecture for Signal Recognition **279**

6.5 APPLICATIONS TO ELECTROMAGNETIC WAVE PROPAGATION IN THE IONOSPHERE **280**

6.5.1 Over-the-Horizon Radar Spectra **280**

6.5.2 Spectral Models **284**

6.6 SUMMARY **284**

6.7 APPENDIX **285**

Notes **286**

Bibliographical Notes **286**

Problems **287**

7 Prediction, Tracking, and Dynamic Models 289

7.1 PREDICTION, ASSOCIATION, AND NONLINEAR REGRESSION **290**

7.1.1 Multidimensional Linear Regression **290**

7.1.2 Multidimensional Autoregression **291**

7.1.3 Nonlinear General Fuzzy Regression ANS (GFRANS) **292**

7.1.4 Nonlinear Autoregression **294**

7.1.5 Example: Data Mining and Revenue Prediction **295**

7.1.6 Summary of Section 7.1 **296**

7.2 ASSOCIATION AND TRACKING USING BAYESIAN MFT **297**

7.2.1 Concurrent Association and Tracking (CAT) **297**

7.2.2 Linear Model for Tracking **299**

7.2.3 Second-Order Model for Tracking **300**

7.2.4 Link-Track Model **300**

7.2.5 Random Noise and Clutter Model **301**

7.2.6 Active Sensor and Doppler Track Models **301**

7.2.7 Autoregression Model for Tracking **302**
7.2.8 Models for Tracking Resolved Objects **303**
7.2.9 Object-Track Declaration **303**

7.3 ASSOCIATION AND TRACKING USING SHANNON–EINSTEINIAN MFT (SE-CAT) **304**
7.3.1 Association and Tracking in Radar Spectral Data **304**
7.3.2 Association and Tracking of Spatiotemporal Patterns **306**
7.3.3 CAT of Spatiotemporal Patterns Described by General PDE Models **307**
7.3.4 Examples of Concurrent Association and Tracking in Radar Data **308**

7.4 SENSOR FUSION MFT **312**
7.4.1 Information Fusion Problem **312**
7.4.2 Mathematical Formulation **312**
7.4.3 Can Sensor Fusion Degrade Performance? **313**

7.5 ATTENTION **314**
Notes **316**
Bibliographical Notes **316**
Problems **316**

8 Quantum Modeling Field Theory (QMFT) 321

8.1 QUANTUM COMPUTING AND QUANTUM PHYSICS NOTATIONS **321**
8.1.1 Quantum vs. Classical Computers **321**
8.1.2 Quantum Physics Notations and the QMF System **322**

8.2 GIBBS QUANTUM MODELING FIELD SYSTEM **324**

8.3 HAMILTONIAN QUANTUM MODELING FIELD SYSTEM **326**
Bibliographical Notes **327**
Problem **328**

9 Fundamental Limitations on Learning 329

9.1 THE CRAMER–RAO BOUND ON SPEED OF LEARNING **329**
9.1.1 CRB, Neural Networks, and Learning **329**
9.1.2 Classical CRB for the Gaussian Means **330**
9.1.3 CR Theorem **331**
9.1.4 CRB for General MLANS Concurrent Association and Estimation **333**

9.2 OVERLAP BETWEEN CLASSES **335**
9.2.1 Overlap Matrix **335**
9.2.2 Overlapping Parts of Means **336**
9.2.3 Overlapping Parts of Covariance Matrices **336**
9.3 CRB FOR MLANS **339**
9.3.1 CRB for Prior Rates **339**
9.3.2 CRB for Means **341**
9.3.3 CRB for Covariances **341**
9.3.4 MLANS Performance vs. CRB: Example 3 Continuation **342**
9.4 CRB FOR CONCURRENT ASSOCIATION AND TRACKING (CAT) **344**
9.4.1 CRB for Linear Tracks **344**
9.4.2 Rule-of-Thumb CRB for CAT **345**
9.5 SUMMARY: CRB FOR INTELLECT AND EVOLUTION? **348**
9.6 APPENDIX: CRB RULE OF THUMB FOR TRACKING **349**
Notes **353**
Bibliographical Notes **354**
Problems **354**

10 Intelligent System Organization: MFT, Genetic Algorithms, and Kant 356

10.1 KANT, MFT, AND INTELLIGENT SYSTEMS **357**
10.1.1 Understanding Is Based on Internal Models **357**
10.1.2 Judgment Is Based on Similarity Measures **359**
10.1.3 Reason Is Based on Similarity Maximization **361**
10.1.4 Hierarchical Organization of Intelligent Systems **361**
10.1.5 Aristotle, Kant, Zadeh, MFT, Anaconda, and Frog **363**
10.2 EMOTIONAL MACHINE (TOWARD MATHEMATICS OF BEAUTY) **366**
10.2.1 Cyberaesthetics or Intellectual Emotions **366**
10.2.2 Purposiveness, Beauty, and Mathematics **367**
10.2.3 Instincts, "Lower Emotions," and Psychological Types **368**
10.3 LEARNING: GENETIC ALGORITHMS, MFT, AND SEMIOSIS **370**
10.3.1 The Origin of A Priori Models **370**
10.3.2 Genetic Algorithms of Structural Evolution **371**
10.3.3 MFT, CAS, and Evolution of Complex Structures **372**
10.3.4 Semiosis: Dynamic Symbol **375**
Notes **378**

Bibliographical Notes **378**
Problems **379**

PART THREE: FUTURISTIC DIRECTIONS: FUN STUFF: MIND—PHYSICS + MATHEMATICS + CONJECTURES

11 Gödel Theorems, Mind, and Machine 383

11.1 PENROSE AND COMPUTABILITY OF MATHEMATICAL UNDERSTANDING **383**

11.2 LOGIC AND MIND **385**

11.3 GÖDEL, TURING, PENROSE, AND PUTNAM **387**

11.4 GÖDEL THEOREM VS. PHYSICS OF MIND **388**
Note **390**
Bibliographical Notes **390**

12 Toward Physics of Consciousness 391

12.1 PHENOMENOLOGY OF CONSCIOUSNESS **392**
12.1.1 Popular Conceptions and Misconceptions about Consciousness **392**
12.1.2 What Is Consciousness? **393**
12.1.3 Consciousness of Bodhisattvas **395**
12.1.4 Consciousness versus Unconscious **396**
12.1.5 Consciousness versus Emotions **397**
12.1.6 Why Is Consciousness Needed? **400**
12.1.7 Collective and Individual Consciousness **400**
12.1.8 Consciousness, Time, and Space **403**
12.1.9 MFT and Searle Revisited **404**
12.1.10 Neural Structures of Consciousness **409**

12.2 PHYSICS OF SPIRITUAL SUBSTANCE: FUTURE DIRECTIONS **412**
12.2.1 Path to Understanding **412**
12.2.2 Physical Nature of Symbol and the Emergence of Consciousness **414**
12.2.3 Nature of Free Will and Creativity **415**
12.2.4 Mysteries of Physics and Consciousness: New Physical Phenomena? **418**

12.3 EPILOGUE **419**
Notes **422**
Bibliographical Notes **423**

LIST OF SYMBOLS 425
DEFINITIONS 429
BIBLIOGRAPHY 447
INDEX 461

This page intentionally left blank

PREFACE

This book describes a new mathematical concept called modeling field theory; demonstrates applications of neural networks based on this theory to a variety of problems; and analyzes relationships among mathematics, computational concepts in neural networks, and concepts of mind in psychology and philosophy. Deep philosophical questions are discussed and related in detail to mathematics and the engineering of intelligence. The book is directed toward a diverse audience of students, teachers, researchers, and engineers working in the areas of neural networks, artificial intelligence, cognitive science, fuzzy systems, pattern recognition and machine/computer vision, data mining, robotics navigation and recognition, target tracking, sensor fusion, spectrum analysis, time series analysis, and financial market forecast. Mathematically inclined philosophers, semioticians, and psychologists will find many issues of interest discussed. Although graduate level is assumed, interested undergraduates will find that most of the material is readily accessible.

Architectures and learning mechanisms of modeling field neural networks utilize a concept of an internal "world" model. The concept of internal models of the mind originated in artificial intelligence and cognitive psychology, but its roots date back to Plato and Aristotle. Intelligent systems based on rules utilize models (rules) in their final conceptual forms. Like the *Eide* (Ideas) of Plato, rules lack adaptivity. In modeling field theory, the adaptive models are similar to the Forms of Aristotle and serve as the basis for learning. By combining the a priori knowledge of models with adaptive learning, the new mathematical concept addresses the most perplexing problems in the field of neural networks and intelligent systems: fast learning and robust generalization. An important aspect of this mathematical and engineering advancement is the discovery of a new type of instinct, a basic instinct to learn, and the role of the related affective signals in general learning. Modeling field theory serves as a stepping stone toward mathematical description of the general phenomena of mind identified by Kant: Understanding (pure reason), Judgment (including higher emotions, beautiful, and sublime) and Will (practical reason and freedom). The combination of intuition with a mathematically unified paradigm provides the foundation of a physical theory of mind.

The book is based on a number of conference presentations and journal publications. It summarizes results of a large research and development effort: during the past 12 years I have been leading a large and successful government-funded neural network program at Nichols Research Corporation. It was expanded to commercial applications, most notably data mining in several areas. In 2000, new commercial companies were formed including InnoVerity (for developing applications in the areas of internet and bioinformatics) and

Ascent Capital Management (for financial predition and investment management). The book describes applications to a number of complicated, real-world problems that have not been solved in the past by other approaches. These applications address pattern and image recognition, data mining, nonlinear time series prediction and spectrum estimation, tracking of patterns in data and imagery sequences, using a variety of sensors and information sources, and the problems of sensor and information fusion.

The first three chapters review mathematical and philosophical concepts of intelligence and mind. Chapter 1, the introduction, begins with the discussion of mathematical approaches to intelligence during the past 50 years and their relationships to philosophical concepts of mind during the 2300 years since Plato. Classical mathematical concepts of hypothesis choice, pattern recognition, prediction, association, tracking, and sensor fusion are reviewed in a concise, mathematically unified framework. This original, unified mathematical framework is presented with an eye toward modeling field theory, which is gradually developed throughout the book.

Chapters 2 and 3 review concepts of the mind in mathematics, engineering, philosophy, psychology, and linguistics and analyze fundamental computational concepts of major algorithmic and neural network paradigms. This analysis provides continuity to a large variety of seemingly disparate techniques and establishes relationships between contemporary computational concepts of modeling intellect and concepts of mind discussed over 2300 years. I found this interrelationship to be much closer than currently thought among scientists and philosophers of today. From the contemporary point of view, the questions about mind posed by ancient philosophers are astonishingly scientific. Contemporary mathematical concepts of intellect are traced as a continuous line through the entire history of psychology and philosophy to the concepts of mind developed in Buddhism, Judaism, Islam, Christianity, and ancient Greece. This interrelationship is emphasized throughout the book. I discuss specific mathematical reasons that lead to a conclusion that knowledge has to be given to us a priori, that is inborn. I show that this knowledge cannot be given as expert rules similar to the Ideas of Plato, but has to be given in a different representation, as in the Aristotelian Forms of mind, which correspond to modeling fields in my theory. The origin of Aristotelian mathematics is traced in Grossberg's ART neural network, in the concept of neural field theory, and in similar concepts of other neural networks. It is a striking conclusion that philosophers of the past have been closer to the computational concepts emerging today than pattern recognition and AI experts of just few years ago. Chapter 2 analyzes learning requirements for each fundamental computational concept and considers relationships between learning requirements, computational complexity, Turing, and physical computability. Chapter 3 relates mathematical and engineering analysis to philosophical analysis. It turns out that fuzzy logic, introduced by Zadeh 2300 years after Aristotelian logic, is an essential ingredient for developing mathematical concepts of the mind based on the Aristotelian theory of Forms.

Chapters 4 through 10 present the new mathematical apparatus for modeling intelligence, with examples of engineering applications. The modeling field theory (MFT) is introduced in Chapter 4. Its three main components are internal models, measures of similarity between the models and the world, and adaptation laws. Deterministic, stochastic, and fuzzy variabilities in data are discussed, followed by an introduction of the concept of modeling fields. A general theory of similarity between a set of models and the world is developed. Aristotelian, fuzzy, and adaptive-fuzzy similarities are considered. Maximization

of adaptive-fuzzy similarity leads to dynamic learning equations of modeling field theory. Two types of adaptive-fuzzy similarity are formulated based on two fundamental concepts of statistics and information theory: Bayesian likelihood and Shannon's information. The principle of similarity maximization, and in particular maximization of likelihood and information, is discussed as an internal drive or instinct to improve the internal representation of the world, that is, an instinct to learn.

Chapters 5 through 7 develop several specific model-based neural networks for various applications of increasing complexity. Chapter 5 discusses the Maximum Likelihood Adaptive Neural System (MLANS), based on Bayesian similarity, for pattern and image recognition applications. Chapter 6 considers Shannon–Einsteinian similarity and discusses Modeling-field Einsteinian ANS (MEANS) for spectrum estimation of transient signals in the frequency domain and in the two-dimensional time-frequency domain. Chapter 7 discusses dynamic temporal and spatiotemporal models for prediction, association, tracking, and recognition of objects and spatiotemporal patterns. Tracking multiple patterns is related to nonlinear time series prediction and tracking applications are discussed along with financial market prediction. Association models are extended to multiple sensor fusion and related to mechanisms of attention. These chapters contain numerous examples of applications to complex real-world problems, many of which could not have been previously solved.

Chapter 8 addresses a possibility that biological neurons may perform quantum computations. A quantum computation algorithm for MFT is described. Chapter 9 considers general limitations on learning for any intelligent system, algorithm, or neural network. Fundamental bounds on learning (the Cramer–Rao bounds) are discussed and new types of bounds are presented for clustering, association, tracking, and nonlinear prediction. Is it possible to compute the fundamental mathematical bound on the entire evolution process?

Chapter 10 discusses the architecture and organization of an intelligent system. The three-component mathematical structure of the modeling field theory is related to the three main components of intelligence identified by Kant: Understanding, Judgment, and Will. Hierarchical and heterarchical organization of Kant–MFT intelligent systems is related to genetic algorithms, complex adaptive systems, and semiotics. A dynamic nature of symbol is discussed. What are the relationships between emotions and thinking? Is a mathematical theory of emotional intellect possible? What kinds of internal models are needed for higher emotional feelings and ethics? Learning behavior leads to improving the internal model, and its mechanisms are related to Kantian reflective judgment—a foundation of higher intellectual abilities. The mathematics of the learning instinct is related to the concept of beauty.

The last two chapters, 11 through 12, contain fun stuff: philosophy and psychology are combined with conjectures based on physical and mathematical intuition about mind. Chapter 11 considers general limitations of logic, computational complexity, Turing computability, and Gödel theorems. Are Gödel theorems relevant to the problems of recognition? Are difficulties encountered by algorithms and neural networks of mathematical intelligence related to Gödel theorems? Does it explain the difference between a human and a machine mind? the nature of free will and creativity? Chapter 12 discusses a possibility of the physical theory of consciousness based on modeling field theory. I discuss the differentiated phenomenology of consciousness and creativity within a framework of modeling field

theory. The Epilogue presents a fresh view on the main discussions of this book: concepts of computational intelligence versus concepts of mind in philosophy, psychology, and linguistics. Can our contemporary mathematical concepts throw light on ancient philosophical problems? Can the thoughts of ancient philosophers guide us in constructing mathematical theories of mind? This book gives affirmative answers to both questions. However, mathematicians and engineers should not be too cavalier about mysteries of the mind, and contemporary philosophers should not bow to mathematical fashions of the day. I attempt to delineate a fuzzy boundary separating questions that today are beyond the scientific method. The book ends with a consideration of the future directions of research in the physical theory of mind.

How to Read This Book

The book is self-contained in that the concepts of philosophy and mathematics are introduced from the basics. Detailed references are provided for further exploration of individual topics. The Definitions section at the end of the book summarizes all the important concepts used throughout the book in alphabetical order for easy reference. The following table provides guidance for several types of readers, who might prefer to read this book selectively.

Concepts	**Chapters and Sections**
General Philosophical Concepts of Intellect and Their Relationships to Mathematical Concepts	Ch. 1, Sec. 1.1 Ch. 2, Sec. 2.1, 2.8 Ch. 3 Ch. 9, Sec. 9.5 Ch. 10 through 12
Overview of Basic Mathematical Concepts of Modeling Intelligence and Their Relationships to Philosophical Concepts	Ch. 1, Sec. 1.2, 1.3, 1.4 Ch. 2 Ch. 9 through 12
Mathematical Concepts and Techniques Related to Specific Applications with Intermittent Discussions of Philosophical Connections	Ch. 1, except Sec. 1.1 Ch. 2, except Sec. 2.1 Ch. 4 through 9

Bibliographical references as well as cross-references among the book chapters were kept to the minimum within the main text. These are contained in Notes and Bibliographical Notes at the end of each chapter, as well as in the Bibliography section at the end of the book.

Semester Courses: Suggested Outlines

Several semester courses can be designed using this book. The following table outlines a few suggestions.

Course Title and Description	Book Chapters
1. Introduction to Modern Pattern Recognition, Prediction, Tracking, and Fusion. A general unified mathematical formulation of problems and solution methods in several areas of statistics and signal processing. Prerequisites: probability Desirable: signal processing Level: graduate or advanced undergraduate	Chapter 1 (Sec. 1.1 is optional), plus any of the examples from Chapters 5 through 7. Or, use your favorite problems.
2. Mathematical Concepts of Intelligence. The course reviews classical mathematical concepts of intelligent algorithms, symbolic AI, and neural networks. After analysis of successes and deficiencies of the classical techniques, new emergent concepts are introduced: evolutionary computation, hierarchical organization, and neural fields. Prerequisites: probability Desirable: a course in neural networks or AI Level: graduate or advanced undergraduate	Chapter 2 (Sec. 2.1 is optional), plus any of the examples from Chapters 5 through 7. Or, use your favorite problems.
3. Model-Based Neural Networks: Statistical Models. Internal models of the world are considered an essential part of intelligence in AI, cognitive sciences, and psychology. The course describes how to design neural networks with internal models. Model-based neural networks combine domain knowledge with learning and adaptivity of neural networks. Prerequisites: probability Level: graduate or advanced undergraduate	Chapter 5.
4. Model-Based Neural Networks: Dynamic Models. Internal models of the world are considered an essential part of intelligence in AI, cognitive sciences, and psychology. The course describes how to design neural networks with internal models. Model-based neural networks combine domain knowledge with learning and adaptivity of neural networks. Prerequisites: probability and signal processing Level: graduate or advanced undergraduate	Chapters 6 and 7.
5. Relationships between Philosophical and Mathematical Concepts of Mind (for students with hard-science background). Relationships between contemporary mathematical concepts of intelligence and 2300-year-old philosophical concepts of mind are much closer than is generally recognized. Specific mathematical concepts and debates are related to specific philosophical ones. Prerequisites: a course in AI, neural networks, pattern recognition, signal processing, or control Level: graduate or undergraduate	Chapter 2, Chapter 3, Sec. 3.1, Chapters 10 through 12.
6. Relationships between Philosophical and Mathematical Concepts of Mind (for students without hard-science background). Relationships between contemporary mathematical concepts of intelligence and 2300-year-old philosophical concepts of mind are much closer than is generally recognized. Specific mathematical concepts and debates are related to specific philosophical ones. Prerequisites: a course in classical or contemporary philosophy Level: graduate or undergraduate	Chapter 2, Chapter 3, Sec. 3.1, Chapters 10 and 11 (with mathematical contents being optional), Chapter 12.

ACKNOWLEDGMENTS

A number of people contributed to this book in various ways: through discussions, encouragement, and support. The ideas within this book were gradually taking shape while, together with my coworkers, we were solving our everyday research and development problems at Nichols Research. My sponsors were interested in these ideas to the extent that they provided financial support for the research, on which this book is founded, and some of them have been actively working on similar ideas. Some read my papers and manuscripts and gave valuable advice. Many issues were clarified in the heated Internet discussions among the subscribers to the Architectures for Intelligent Control Systems list. My friends attending Friday night gatherings at Jordan road provided much discussion and thought. And my wife inspired me to think profoundly.

It is my pleasure to thank the people whose thoughts, ideas, encouragement, and support shaped this book and made it possible: M. Akhundov, J. Albus, U. Aleshkovsky, R. Berg, R. Brockett, B. Burdick, G. Carpenter, W. Chang, D. Choi, R. Deming, V. Dmitriev, W. Freeman, K. Fukunaga, L. Garvin, R. Gudwin, M. Gouzie, S. Greineder, S. Grossberg, M. Karpovsky, M. Kreps, L. Levitin, A. Lieberman, T. Luginbuhl, A. Meystel, K. Moore, V. Oytser, D. Radyshevsky, C. Plum, A. Samarov, W. Schoendorf, D. Skatrud, R. Streit, E. Taborsky, E. Tichovolsky, B. Veijers, D. Vinkovetsky, Y. Vinkovetsky, V. Webb, M. Xiarhos, L. Zadeh, and G. Zainiev.

Neural Networks and Intellect

This page intentionally left blank

part one

OVERVIEW

2300 Years of Philosophy, 100 Years of Mathematical Logic, and 50 Years of Computational Intelligence

This part of the book consists of three chapters: Chapter 1 is an introduction to the book and to the concepts of intelligence in philosophy and mathematics. Chapter 2 reviews mathematical concepts of intelligence. And Chapter 3 relates the mathematical concepts to the philosophical concepts of intelligence from Plato to Jung.

This page intentionally left blank

INTRODUCTION
Concepts of Intelligence

This chapter serves as an introduction to the book. It begins with an overview of the history of mathematical and philosophical concepts of intellect illustrating close relationships between the two areas. Then, it reviews classical mathematical concepts used in the design of intelligent algorithms: theory of probability, Bayesian hypothesis testing, pattern recognition, estimation, clustering, association, and prediction methods for a single process (regression and autoregression) and for multiple processes (tracking). These diverse areas are reviewed in a concise, mathematically unified framework. The original, unified mathematical framework serves as an introduction to modeling field theory, which is gradually developed throughout the book. We review concepts of intelligent systems' architecture and organization. A mathematical concept of the internal model is introduced. Its fundamental role in intelligence is established by relating it to the philosophical concepts of mind.

1.1 Concepts of Intelligence in Mathematics, Psychology, and Philosophy

This section overviews the history of concepts of intellect and serves as an introduction to metaphysical and mathematical analysis in Chapters 2 and 3.

1.1.1 What Is Intelligence?

The human mind, intelligence and its limits, the range of spiritual human experiences and computers, artificial intelligence, robots, the Internet's sea of information, and as yet unexhausted possibilities—what a huge area for study and research! But how often we have to say to ourselves in despair: is it possible to explore these vast spaces? Thousands of powerful minds have treaded here. Is it possible to grasp the expanses of their thoughts? And, to go beyond? But curiosity has its rewards in heaven and even on earth.

What is the subject of this book? Is there a definition of intelligence? Do we need one? In my opinion, clear definitions appear at the end of research, so I will not worry about the absence of a concise definition at the beginning. Notwithstanding, as a first step

and a suggestion for further thinking, let us characterize intelligence as a goal-directed functioning. Then, one may add functioning inside and outside of an intelligent system self; selection of goals and subgoals; sensing, perception, recognition, decision making, planning, acting; acting inside and outside of the self; learning and adaptation; memory; acquiring, storing, and using knowledge; hierarchical and parallel organization (of all of the above: goals, functioning, knowledge); reproduction; evolution; social organization; organization of environment; organization of self. This list should be continued toward thinking, feeling, emotion, intuition, consciousness, free will, and creativity.

But then, where shall we start? What is the minimal subset of the above properties that would lead to an interesting, nontrivial theory of intelligence? The theory, on the one hand, should be useful in engineering applications, and on the other should conform to our knowledge of psychology, neurobiology, brain organization, and, if not satisfy, at least not offend too much our intuition of what is intelligence, or what I call a physical intuition about spiritual substance. This first section of this chapter can be considered as an introduction to the topic. I will show that a particular relatively small subset of the above properties attracted significant attention in our mathematical research of intelligence during the past 50 years and spurred philosophical debates during the past 2300 years. This fascinating property of intelligence is an ability to combine a priori knowledge (available before experience) and adaptive learning (from experience). Over more than two millennia, these properties seemed to symbolize the basic aspects of mind, while at the same time they seemed mysterious to philosophers and elusive to mathematicians. Let us review the history of debates of apriority and adaptivity of mind in mathematics and philosophy.

1.1.2 Plato, Occam, and Neural Networks

A contemporary direction in the theory of intellect is based on modeling neural structures of the brain. It was founded by McCulloch and co-workers beginning in the early 1940s. McCulloch intended to create a mathematical theory of intellect on the basis of complicated a priori neural structures. The basis of this search for the material structures of intellect, for the explanation of how the interactions of brain neurons could process the information and perform computations was founded on a *realistic* philosophy, created by the school of Plato and Aristotle.

Plato, 2300 years ago, came to a conclusion that the ability to think is founded in the a priori knowledge of concepts: the concepts or abstract ideas (*Eide*) are known to us a priori, through a mystic connection with a world of Ideas. For example, *chair* as a concept describes the whole class of objects (individual chairs). The Eide, according to Plato, have true existence or are real in some sense in which our everyday concrete experience lacks reality. This conception that, at first glance, might seem ridiculous to a scientific mind has been much debated throughout antiquity and the Middle Ages, and is being debated today, unexpectedly turning into the basis of many algorithms of artificial intelligence. Aristotle, Plato's pupil, critiqued his teacher's theory by pointing out that it does not account for an important aspect of the intellect—an ability to *learn or adapt* to a changing world. Throughout early antiquity and the Middle Ages, concepts of Plato and Aristotle were unified into a grand philosophical system based on the *realism* of Ideas. The ways in which the *intellect combines apriority with adaptivity*, and is determined by the measured play of

these two factors, has remained at the center of philosophical, theological, and mathematical debates on the nature of mind.

According to McCulloch, the a priori Eide of Plato were encoded in the complicated neural structures of brain. In search of a mathematical theory unifying neural and cognitive processes, McCulloch and co-workers combined an empirical analysis of biological neural networks with information theory and mathematically formulated important properties of neurons. McCulloch and Pitts (1943) reduced a complicated entanglement of a large number of complex factors characterizing biological neurons to a few important properties necessary for mathematical modeling of the neural organization of the brain. They created a simple mathematical model that was later named the formal neuron. This model was supplemented by an adaptation mechanism by Hebb in 1949, and it served as a basis for creation of the first artificial neural networks. The first neural network utilizing properties of formal neurons was built by Minsky and Edmonds in 1951 using tubes, motors, and clutches, and it modeled the behavior of a rat searching for food in a maze.

In the 1950s, neural networks utilizing formal neurons were developed by several groups of researchers including Rosenblatt and Widrow. Widrow's adalines utilized a cybernetic concept of control based on simple models, Wiener filters, that led to fast learning in linear signal filtering problems (Widrow, 1959). Perceptrons created by Rosenblatt (1958) were capable of learning linear classification rules from training data. Thus, perceptrons learned classes of similar input data patterns, or in other words, they learned "concepts" from empirical data! Early neural networks utilized simple structures. It was expected that a large number of adaptive neurons connected into a network would be able to learn complex cognitive and behavioral concepts on their own. A priori knowledge, it seemed, was not needed, and could not be utilized by early neural networks. The complex neural structures postulated by McCulloch were not needed nor was the reality of the Plato's Eide: *concepts* could be learned from experience.

This view on the origin of concepts of mind was not new. Occam, who lived in the fourteenth century and is considered one of the last great medieval scholastic thinkers, rejected the realism of Plato and Aristotle. He felt that predominantly theological thinking emphasizing the a priori aspect of intellect based on God-given knowledge had a stifling influence on the development of knowledge. Following Antisthenes, founder of the Cynic school of philosophy, Occam held nominalistic views. *Nominalism* considers ideas to be just names (nomina) for classes or collections of similar empirical facts. For example, a concept *chair* is just a name for the class of objects (individual chairs). Nominalism emphasizes the ability of mind to learn from experience. Occam set to overcome the limiting influence of the conception of apriority. He came to believe that only particular experiences have real existence and that general concepts (universals) are just names for similar types of experiences, devoid of any real existence. Analyzing the empirical, experiential origin of knowledge, Occam developed the basis for the coming philosophy of empiricism, which was essential for the development of the scientific method in the following centuries. His work indicated (or initiated?) a shift of interest *away* from spiritual, mental processes, *away* from the question of the rational understanding of the intellect, and toward an objectified method of inquiry, which later became associated with the scientific method.

McCulloch believed that the nominalistic way of thinking was detrimental to the development of theories of mind: "under the influence of nominalistic concepts since Occam, the realistic logic decayed, which caused problems for scientific understanding of

mind." The attempt by McCulloch to found a new theory of neural networks on a *realistic* philosophy was revolutionary and counter to the 500 year evolution of the mainstream of the scientific method. This revolutionary attempt to understand the mind based on the apriority of concepts was short lived. As mentioned, early neural networks deviated from the program outlined by McCulloch: their learning was based entirely on experience, unaided by specialized a priori structures.

The early research in neural networks from the 1940s to 1960s generated tremendous interest as it promised to resolve the mystery of the mind. Why did the Goliath-to-be fall down in the late 1960s? How did it happen that a relatively mild criticism of perceptrons by Minsky and Papert in 1969 had a devastating effect on the interest in artificial neural systems? The question of why this happened was widely discussed in the scientific community. However, the often offered explanations pointing to personal opinions cannot be accepted, since they are unscientific and relatively useless. A personal opinion can produce a large scale effect in a society only if it captures, embodies, and serves as a conduit for a changing philosophical trend. The crisis in the field of early neural networks coincided with the contemporaneous downfall of behavioristic psychology and philosophy that share nominalistic origins. Simple structures of early neural networks and learning based entirely on concrete empirical data were in agreement with the nominalistic concept of the intellect dominant at the time. This begs a question: Was this association not the real, philosophical reason for the downfall of the early neural network research—brought about by the downfall of behaviorism, a philosophy no longer tenable—rather than by scientific criticism? It seems that scientific and mathematical paradigms are directly related to the philosophical debates of the past and to shifts in metaphysical paradigms of thought between analytic and holistic, spiritual and material, empirical and a priori. Thus, it is revealing to trace the metaphysical origins of our mathematical concepts of intellect.

1.1.3 Rule-Based Artificial Intelligence, Complexity, and Aristotle

Near the end of the 1960s, being dissatisfied with the existing capabilities of mathematical methods of modeling neural networks, Minsky suggested a different concept of artificial intelligence that descended from Plato's principle of apriority of ideas. For a computer to operate and make decisions in a complicated environment, concluded Minsky, knowledge ought to be placed into the computer a priori. In Minsky's method, named expert or rule systems, a system of logical rules is put into a computer. This system contains all possible situations (for example, all possible readings of sensors of a particular device or system) and expert decisions or rules of what is to be done in each particular situation. This method, which I will call the Plato–Minsky approach,[1] became the foundation for many practical applications of computers, from factory floors to space shuttles. It was the next attempt (after McCulloch) to understand the intellect on the principle of realism of ideas.

Answering the very first question of intelligence: How is intelligence possible?—the Plato–Minsky approach does not explain an important aspect of mind—an ability to learn and to adapt, leaving unanswered the second question about intelligence: How is learning possible?

Although in 1975 Minsky emphasized that his method does not solve the problem of learning, notwithstanding, attempts to add learning to Minsky's artificial intelligence have

been continuing in various fields of modeling the mind, including linguistics and pattern recognition throughout the 1970s, 1980s, and continue today. In linguistics, Chomsky proposed to build a self-learning system that could learn a language similarly to a human, using a symbolic mathematics of rule systems. In Chomsky's approach, the learning of a language is based on a language faculty, which is a genetically inherited component of the mind, containing an a priori knowledge of language. This direction in linguistics, named the Chomskyan revolution, was about recognizing the two questions about the intellect (first, how is it possible? and second, how is learning possible?) as the center of a linguistic inquiry and of a mathematical theory of mind. However, combining adaptive learning with a priori knowledge proved difficult: variabilities in data required more and more detailed rules leading to combinatorial complexity of logical inference. Combinatorially complex solutions are not physically realizable for complicated real-world problems.

Here, we just met with a ubiquitous problem of combinatorial complexity. On the one hand, intelligence should be flexible enough to manipulate various combinations of multiple elementary notions, concepts, and plans in order to find suitable decisions in complex situations. On the other hand, a straightforward evaluation of combinations leads to a combinatorial explosion: we will see that the number of combinations, even for problems of moderate complexity, is very large, exceeding the number of particles in the universe. Therefore, brute-force solutions are impossible. We will be returning to this problem throughout the book.

Concurrently with early neural networks and rule-based intelligence, wide use of digital computers beginning from the 1960s resulted in a large body of self-learning, adaptive algorithms for pattern recognition based on statistical techniques. To recognize objects (patterns of data) using these methods, the objects are characterized by a set of classification features that is designed based on a preliminary analysis of a problem and thus contains the a priori information needed for a solution of this type of problem. Within the limits of similar type problems, these algorithms can adapt by using adaptive statistical models. However, their application to complicated real-world problems that are not limited to a single well-determined type is rarely achievable, because general mathematical methods for the design of classification features have not been developed, and their design based on a priori knowledge remains an art requiring human participation. When problem complexity is not reduced to a few classification features in a preliminary analysis, these approaches lead to difficulties related to exorbitant training requirements. In fact, training requirements for these paradigms are often combinatorial in terms of the problem complexity. These algorithms, therefore, are not suitable as physically realizable models of intellect.

A striking fact is that the first one who pointed out that learning cannot be achieved in Plato's theory of mind was Aristotle. Aristotle recognized that in Plato's formulation there could be no learning, since Ideas (or concepts) are given a priori in their final form. Thus, learning is not needed and is impossible, and the world of ideas is completely separated from the world of experience. Searching to unite the two worlds and to understand learning, Aristotle developed a concept of Form, having an a priori universal reality and being a formative principle in individual experience. In Aristotelian theory of Form, the adaptivity of the mind was due to a meeting between the a priori Form and matter. The major point of Aristotelian criticism of Plato's Eide concepts was that before a Form meets matter, it does not attain its final form of a concept. This theory was further developed by Avicenna (XI), Maimonides (1190), Aquinas (XIII), and Kant (1781) among many other philosophers

during the past 2300 years. Aristotelian Forms are dynamic entities afforded variable degrees of uncertainty before their potentialities are realized. However, Aristotelian logic described laws governing eternal truths, not fluid Forms. For example, the Aristotelian law of excluded third states that every concept (or statement) is either true or false, anything else is excluded. It is more applicable to Plato's Ideas than to Aristotelian Forms.

The contradiction between Aristotelian theory of mind and Aristotelian logic is inherited by contemporary mathematical theories of intellect. Algorithms that are most widely utilized today to combine adaptivity and apriority are based on Aristotelian logic, which is inadequate for this purpose. These algorithms face combinatorial computational complexity and are not suited for real world problems. And a National Science Foundation report concluded that "much of our current models and methodologies do not seem to scale out of limited 'toy' domains."

A mathematical description of the Aristotelian theory of mind should overcome the inadequacy of the Aristotelian logic for this purpose, it should address the a priori Forms and the process of meeting between Forms and matter. A first step toward this was the development of fuzzy logic by Zadeh. Fuzzy logic operates without the law of excluded third; it accounts for the inherent approximate nature of thoughts and concepts. A second step toward mathematics of the Aristotelian theory of mind was made by Grossberg, a founder of contemporary neural network theory. In the 1980s, Grossberg established that a fundamental mechanism of perception and cognition is interaction between signals coming from within the mind and from the outside world (efferent and afferent signals). This is the Aristotelian meeting of Form and matter. It was a fundamental departure from early neural networks, which emphasized learning from data (signals coming from the outside). And it was contrary to the rule-based artificial intelligence that emphasized the role of signals coming from within the mind. A third step that combines (1) fuzzy logic and (2) interaction of efferent and afferent signals with (3) adaptive fuzzy models of the a priori forms is a subject of this book.

1.1.4 Philosophy vs. Architecture of Intelligent Tracker

Let us relate concepts discussed above to a concrete example of an engineering design. This section describes an intelligent system and relates engineering and mathematical concepts to the philosophical ones using this concrete example. From an engineering standpoint, it is a large-scale complex operational system involving radars and computers, and the description here will be limited to most important concepts related to intelligence. From the point of view of general intelligence, it is a very simple system, comparative to the human or even animal mind. Nevertheless, this example gives us a chance for a concrete discussion of many of the concepts of intelligence discussed above: apriority and adaptivity, learning and combinatorial complexity, concepts, and objects. Even more, we will introduce several new concepts related to intelligence: hierarchical vs. heterarchical organization, internal models, similarity measures, intelligent agents, the nature of signs and symbols, and their relationships to concepts and internal models. This section previews many of the issues that will be discussed throughout the book. We will barely touch on many complex issues here; these issues will be discussed in details later. Therefore, at first reading, I'll suggest that readers skip anything that may seem insufficiently explained or superficial; it might be useful to refer back to this section while reading the rest of the book.

The description below consists of two parts: (1) an overall architecture and (2) a more detailed discussion of the architecture at the most "interesting" level.

1.1.4.1 A Hierarchical Architecture

This tracker is a hierarchical system comprised of several layers. It is designed to detect and track aircraft and ships within an area of several million square miles. At the bottom level, there are the radar data (signal strength), dimensioned by range, azimuth, doppler velocity, and time; this can be envisioned as a time sequence of three-dimensional images.

At the next level there are two types of intelligent agents: search agents and track agents. Each agent is an automaton or a semiautonomous subsystem. The entire field of data over some time window is "watched" by about a hundred of search agents, which are looking for track-like events. Each agent is responsible for his "territory" (about 10,000 data pixels). When it "sees" a track-like event, it starts tracking it and becomes a track agent; a new search agent is put in its place by a higher level.

The next level decides which of the track agents are "good" and which do not really track anything.

The next level forms long-duration tracks.

The next level makes corrections to these tracks (there is a large number of things related to the complex nature of the propagation of the radiowaves through the ionosphere).

The next level interacts with operators: displays results and accepts operator corrections.

The next level interacts with the user of the system: a high-level military commander.

1.1.4.2 "Interesting" Architectural Details: Intelligent Agents

An architecture of the search and track agents implements several concepts of general intelligence. Each search or track agent has three subsystems: (1) internal model (IM), (2) similarity measure or association subsystem (AS), and (3) adaptation law or parameter estimation subsystem (PS). The agent operations consist in iterative performance of $1 \rightarrow 2 \rightarrow 3 \rightarrow 1 \ldots$ This iteration always converges (that is, after few iterations, parameters reach their proper values and do not change much thereafter).

1. IM is a parametric model of an object-track (the law of motion plus the law of radar signal propagation and scattering); its parameters are the track state vector (position, velocity, radar cross section) and its errors. From these parameters, IM computes the (expected or predicted) position of the track, its expected errors, and the radar signal strength. Note that here we are talking about two different levels of the model representation: the concept-model (laws and parameters or attributes) and object-model (computed expected signal).
2. AS computes similarity measures between each pixel in the agent's field of view and the computed track signal (it associates track with data). This computation accounts for the expected track errors computed above. It also computes the overall similarity measure between its track and all its pixels, which is used by the higher level to decide on continuation or killing of the track-agent.
3. PS estimates the IM parameters (the track state vectors and their errors). This estimation is based on the association computed in step (2).

1.1.4.3 Comments: Philosophy vs. Mathematics

The IM contains the a priori knowledge: its parametric form is given a priori. The IM is an adaptive model: its parameters are computed adaptively, from data; as data change, the parameters may change as needed. Thus, the intelligent tracker combines the apriority and adaptivity.

The IM is a fuzzy model: it is characterized not only by parameters, but also by their expected errors. The errors are used in AS to compute the similarities leading to fuzzy association between pixels and models. The search agent starts with large errors or large fuzziness (therefore, its initial parameter values are not too important). In the process of iterations, errors get reduced and the estimated track parameter values converge to the true values. Thus a search-agent smoothly becomes a track-agent; searching and tracking are different states of the same automaton. This adaptation of the parameters comprises the agent's learning process. The degree of fuzziness is reduced in the process of learning.

This intelligent tracker is different from other approaches to tracking in a fundamental way: it is inherently noncombinatorial. It is commonly believed that complex tracking problems are inherently combinatorial, for the following reasons. To estimate parameters of a track model, it is necessary to know which pixels belong to the track. Therefore, classical approaches to tracking involve first, generating a large, combinatorial number of alternative candidate tracks defined by various combinations of pixels, and second, evaluating which of these tracks are "more likely" according to some criterion. Contrary to this, the intelligent tracker system requires no combinatorial searches. Combinatorial searches are eliminated by fuzzy associations.

Let us analyze the above discussion of the combinatorial problem and its solution in more detail. In classical approaches, alternative candidate tracks are generated according to the Aristotelian logic: a particular pixel either belongs to a track or does not (the third is excluded). This leads to a combinatorial explosion of the number of possible alternative tracks. In the intelligent tracker a single search-agent is associated with all pixels (in its field) in a fuzzy way, excluding a need for the combinatorial search. Thus, fuzzy logic is used to overcome the combinatorial complexity of the Aristotelian logic.

The above analysis is not limited to a tracking problem, but is of a general nature. Many complex problems of recognition, planning, etc., are solved by a structural combination of "primitives" or agents, each solving a small part of the problem. Finding a good structural combination is widely believed to require combinatorial searches, for the same reason as above: the Aristotelian logic used in the search process is inherently combinatorial. Fuzzy logic can be used to overcome this difficulty. To accomplish this, in the general case, as in the case of the intelligent tracker, we need to develop suitable measures of similarity and the adaptation procedures. In other words, we need model-based adaptive fuzzy logic. The development of this technique is one of the main themes of this book.

The process of adaptation of the intelligent tracker resembles the process of learning as described by Aristotle. A highly fuzzy search-agent corresponds to an a priori Form. The process of adaptation to the data corresponds to the meeting of Form and matter: in this process an a priori Form (search-agent) is transformed into a nonfuzzy concept (track-agent).

1.1.4.4 Intelligent Tracker vs. Intelligence

Each intelligent search-track-agent of the tracker possesses a formidable degree of intelligence: an a priori knowledge of a general concept of a track, an ability to recognize

a specific subset in the data that corresponds to this general concept, and an ability to learn a specific concept-object of a particular track. Throughout the book I will argue that these properties represent an essential element of the thought process. And, developing mathematical methods suitable for these type of intelligent agents occupies a significant part of the book.

Compared to the human mind, the agents are not very intelligent. They do not possess much understanding of what they are doing. (One may argue that we, humans, also do not understand much of what we are doing. Still, we understand something.) An agent cannot even be said to understand the meaning of the single concept that it comes up with, the concept-object of a track. At most, we can say that an agent understands the unordered unstructured manifold world around him in terms of this concept of a track. But an understanding of what the track is belongs to the higher level of the architecture of the tracker. There, in the next level, tracks found by various agents are compared to each other, real tracks are sorted out from track-like clutter events, long-duration tracks are formed, and appropriate signal-reports are sent to a next higher level. Establishing these relationships among various concepts is the essence of the understanding of the meaning of these concepts.

At an appropriately high level, characteristics of tracks are compared to the goals that the system is tasked to perform. This comprises an essential element of the understanding of the situation. Based on this understanding, reports are issued to the human commander, which eventually might affect events in the world. The tracker is not capable of generating its own system-level goals (and we are not sure that we, humans, can generate our top-level goals either). Still, the tracker can propagate his human-given system-level goal down to the lower levels and to generate their "subgoals." An example of a subgoal could be to find slow-moving objects, which will be translated into a subsubgoal to allocate more resources to slow-moving objects, etc. An individual agent does not generate behavior in the outer world. Still, there are two types of behavior that each agent performs. Upon convergence, it sends a signal to the higher level. And it performs adaptation of its model to the data, or in other words, it improves its knowledge about the world. Possibly, this latter ability forms the foundation for all or many of our higher intellectual abilities, this will be discussed in Chapter 10.

1.1.4.5 Signs, Symbols, and Tracks

Signs and symbols are essential aspects of intelligence. The nature of signs and symbols and their roles in intelligence are studied by semiotics. This places semiotics close to both mathematics and the philosophy of intelligence. A reader not interested in the nature of signs and symbols can omit this subsection on first reading. Here, I relate the above discussions to semiotical concepts and terminology. For example, consider the following material entity in the world: a written sequence of characters, say "chair." It can be interpreted by a mind to refer to something else: another entity in the world, a specific chair, or the concept "chair" in your mind. In this process, a mind, or an intelligent system, is called *an interpreter*, the written word is called *a sign*, the real-world chair is called *a designatum,* and the concept in the interpreter's mind, the internal representation of the results of interpretation, is called *an interpretant* of the sign. The essence of a sign is that it can be interpreted by an interpreter to refer to something else, a designatum. This is achieved through the interpretant, which, in turn, becomes a sign for the next layer of the interpreter's architecture, where it would

be interpreted as referring to something else, say to the "behavior of sitting in the chair," etc. A collection of the multiple relationships of the interpretant to other concepts refers to the designatum, an object-chair in the world. This is a simplified description of a thinking process, called semiosis. And even this simplified description implies specific consequences for an architecture of any intelligent system.

Note that one of the functions of the intelligent system is to "interpret" the world, that is to develop internal representations of the world and to establish the correspondence between the world and the interpreter's representations. Any structure or object in the world exists only as a result of interaction between the world and interpreter. Establishing structures requires *a measure of similarity,* which has to be represented inside of an intelligent (semiotical) system.

Let us analyze our tracker using the semiotical terminology. A search-track-agent interacts with the world. In the process of this interaction it finds/imposes a structure on the world. It does so by possessing an inherent a priori measure of similarity between its model-track and the world. In the result it comes up with an object-concept: a track of a moving object (track-agent). A semiotical analysis distinguishes the material entity in the world from the sensory data about the object and from the concept in "the mind" of the intelligent system—tracker. The identified structure in the sensory data is *a sign.* It refers to the material entity in the world (a moving object), *a designatum*. The tracker is *an interpreter*. The interpreted sign is represented inside the interpreter by *an interpretant:* a track-agent, which is a concept of a moving object, or a concept-object.

Note a difference between search-agents and track-agents. Search-agents are highly fuzzy and highly adaptive: due to their large error and uncertainty, they can find many different tracks within their field of view. Track-agents are little fuzzy and little adaptive: their errors and uncertainty are small, and if track parameters change drastically, the track can be lost. A search-agent is a process of an emergent concept, whereas a track-agent is a well-established fairly specific concept-object. The dynamic process of formation of an emergent concept out of uncertainty I call a symbol or symbol-process.[2] Search-track-agent is a symbol. It is a dynamic process, a subprocess of the semiosis performed within the entire system. Upon convergence of the search-agent's iterative estimation process, it becomes a track-agent and it sends a signal to the higher levels of the hierarchy. This signal is the interpretant of the moving object. For the higher levels of the hierarchy, this signal is not a dynamic symbol but a simple nonadaptive sign of a track. Note similarities and differences between the process of semiosis and Aristotelian description of a meeting between Form and matter (Problem 1.1-1).

1.1.5 Summary

The relationship between the mathematical and philosophical concepts of mind, touched upon in this section, continues in Chapter 3 and in a less conspicuous way penetrates the entire book. Philosophical analysis helps to place the right emphases in the mathematical analysis and vice versa. Throughout the history of philosophy, concepts of apriority and adaptivity of mind remained in the center of debates, and often split the philosophical community. But the great unifiers of philosophy worked toward combining both factors. The philosophical analysis emphasizes that factors of apriority and adaptivity ought to be combined by physically acceptable concepts of the intellect. This conclusion is compared

with the mathematical analysis of approaches to the design of systems and algorithms of intelligence in Chapter 2. The mathematical analysis leads to a conclusion that there are few basic computational concepts forming the foundation for all the multiplicity of learning algorithms and neural networks. These basic concepts are closely related to the philosophical conceptions of mind, the apriority and adaptivity. Both types of algorithms faced combinatorial explosion, those associated with apriority faced logical complexity, and those associated with adaptivity faced training complexity. Attempts to combine the two led to combinatorial complexity of computations.

The difficulty of combining apriority and adaptivity was traced to the original contradiction in Aristotelian teachings, and the Aristotelian logic was identified as a culprit. This analysis continues in the following chapters. Chapters 2 and 4 discuss a need to use fuzzy logic for the mathematical description of Aristotelian Forms. Thus, fuzzy logic, formulated by Zadeh in 1965, 2300 years after Aristotle, provides the foundation for developing mathematical theory of Aristotelian Forms. The Forms are described in contemporary terms as model-based adaptive fuzzy concepts. The mathematics of Aristotelian theory of mind should combine fuzzy logic with apriority and adaptivity. Such a mathematical theory is developed in Chapter 4, which describes a theory of neural modeling fields combining a priori knowledge with learning and fuzzy logic as a step toward physically acceptable concepts of intellect.

But before turning to these, we need to review several topics of classical theory of probability and statistics. A reader familiar with this subject may skip it, or may choose to look briefly through the rest of this chapter in order to note the notations: systematic notations are introduced here that are used throughout the book, across several areas of statistics and signal processing.

1.2 Probability, Hypothesis Choice, Pattern Recognition, and Complexity

This section introduces classical mathematical concepts and definitions, and relates the main subject of this book, the concepts of intellect, to classical areas of probability, choice of hypothesis, pattern recognition, estimation theory, and prediction. It serves as an introduction to Chapters 3 and 4.

1.2.1 Prerequisite: Basic Notions of the Theory of Probability

Readers familiar with probability theory can skip this section. Here, in simple form, we briefly overview the basic notions, definitions, and notations of probability theory used throughout this book. I emphasize the rationale for the concepts, while keeping the mathematical rigor at the bare minimum, although all of the notions and definitions can be made mathematically rigorous.

Probability theory is used for mathematical modeling of uncertainty. Probability theory begins with a notion of a random variable. A random variable x (or event) can be observed multiple times, and from observation to observation, x varies randomly and does not vary deterministically. For example, x is a particular dice throw, or a card hand, or the result

of a measurement or observation conducted with finite precision. The notion of a random variable is approximate only in practical applications and it does not cover all types of uncertainty (Fig. 1.2-1). When the assumption of randomness does not seem appropriate, one can attempt to separate predictable deterministic effects from random, unpredictable effects. When uncertainty involves unknown circumstances that are not of random origin, one can attempt to use fuzzy variables or fuzzy logic. Advanced methods combining random probabilistic variabilities with deterministic variabilities and with non-random uncertainties are considered later, throughout the book.

Definition of Probability. Probability of an event x, $P(x)$, is the relative frequency of observing event x among all other events (in the limit of infinite number of observations). Thus, if event x is observed N_x times and the total number of observations is N, $P(x) = N_x/N$, in the limit of $N \to \infty$.

For example, a measurement is performed resulting in value x. We call it an event x. Probability $P(x)$ is the relative frequency of measuring this value among all other values of x. If x is a continuous variable, we usually talk about the probability of observing values in a small interval from x to $x + dx$, $P(x \text{ to } x + dx) = f(x)dx$. The function f is called a probability density function (pdf), or simply probability density, and we usually denote it as $\text{pdf}(x)$ (see Fig. 1.2-2).

From the definition, it follows that

$$\sum_x P(x) = 1 \tag{1.2-1a}$$

For example, in a fair coin toss, $P(\text{head}) = P(\text{tail}) = 0.5$, $P(\text{head}) + P(\text{tail}) = 1$. For a continuous variable x, the sum above is substituted by an integral,

$$\sum_x P(x) \to \sum_x \text{pdf}(x)dx \to \int \text{pdf}(x)dx = 1 \tag{1.2-1b}$$

Definition of Independent Events. Two events x and y are called independent if the probability of x is unaffected by occurrence of y and vice versa.

Rule of Combining Independent Probabilities. If x and y are independent events, the probability of the joint event (x, y) is given by $P(x, y) = P(x)P(y)$. Correspondingly, $\text{pdf}(x, y) = \text{pdf}(x)\text{pdf}(y)$ (see Fig. 1.2-2c, and d).

For example, in a fair coin toss, the outcomes of the first and second tosses are independent; the probability of the event of two heads in two tosses $P(\text{head,head}) = P(\text{head})$ $P(\text{head}) = 0.5^2 = 0.25$.

Definition of Conditional Probability. When considering the probability of x, given that event y has occurred, the conditional probability of x given y is denoted $P(x|y)$. For continuous variables, $\text{pdf}(x|y)$ denotes pdf of x given y [that is $P(x \text{ to } x + dx)$ given y]. For independent events, $P(x|y) = P(x)$ (see Fig. 1.2-2d).

Rule of Conditional Probability. Probability of the occurrence of both events, x and y,

$$P(x, y) = P(x|y)P(y), \quad \text{pdf}(x, y) = \text{pdf}(x|y)\text{pdf}(y) \tag{1.2-2}$$

Definition of Alternative Events. Two events x and y are called alternatives if either one can occur with some probability, but not both (see Fig. 1.2-2e).

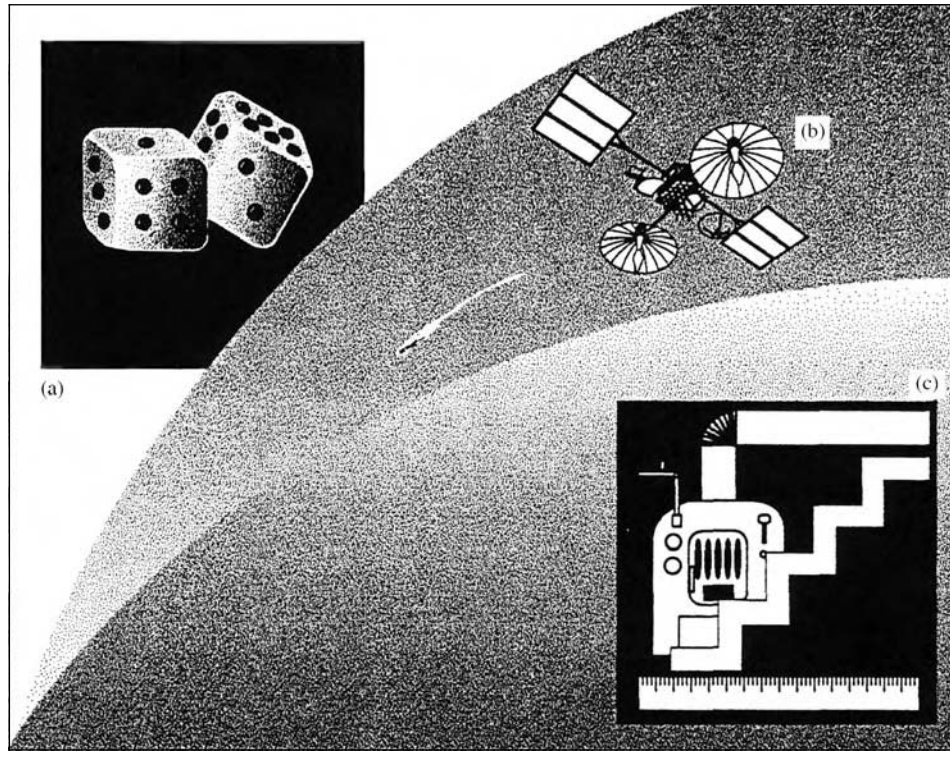


Figure 1.2-1 Examples of probabilistic (a), deterministic (b), and fuzzy (c) events. **(a)** Probability works well for predicting dice throws, even if dice are uneven; **(b)** prediction of satellite motion in orbit is best accomplished by using deterministic Newtonian laws (this is true up to a point; collisions with meteors could be described as random probabilistic events); **(c)** if you never measured the size of your cellar, you may be uncertain about exact dimensions, but if it is not random, fuzzy variables can be appropriate for describing your knowledge.

For example, if we perform one measurement of x, different x values are alternatives.

Definition of a Complete Set of Events. If a set of events exhausts all possibilities, it is called complete (see Fig. 1.2-2f).

Rule of Combining Alternative Probabilities. If x and y are alternative events, the probability of either of the events x or y is given by $P(x \text{ or } y) = P(x) + P(y)$.

For example, x and y could be two different outcomes of the same process, say a single coin toss, $P(\text{head or tail}) = P(\text{head}) + P(\text{tail})$. For a complete set of alternatives, the sum of probabilities is 1. If the coin is unfair, $P(\text{head}) \neq P(\text{tail})$, but still, $P(\text{head}) + P(\text{tail}) = 1$. Another example: Eq. (1.2-1): all different values of x form a complete set of alternatives.

Consider a more complicated combination of alternatives. For example, tomorrow the Federal Reserve Board is going to announce its decision concerning interest rates, and your subjective probabilities for their actions are (1) interest rates will go down with probability $P(1) = 0.3$, (2) interest rates will not change with probability $P(2) = 0.5$, (3) the decision will be postponed with probability $P(3) = 0.2$, or (4) interest rates will go

up with probability $P(4) = 0.0$. Only one of these four can actually occur, therefore, these are the four alternative processes or events that might affect your portfolio. You would like to predict how this will affect your portfolio, and you have a model that tells you that in case (1), the probability density for the price of your portfolio x is $\text{pdf}(x|1)$ and correspondingly, you have $\text{pdf}(x|2)$, $\text{pdf}(x|3)$, and $\text{pdf}(x|4)$. In this case, the total pdf for tomorrow's price of your portfolio, taking into account the four alternatives and the corresponding models, is given by

$$\text{pdf}(x) = P(1)\text{pdf}(x|1) + P(2)\text{pdf}(x|2) + P(3)\text{pdf}(x|3) + P(4)\text{pdf}(x|4) \quad (1.2\text{-}3)$$

This equation is a consequence of the two rules: conditional probabilities and combining alternatives. Note, when writing the above equation it is important to account for *all* alternatives, that is *a set of alternatives has to be exhaustive*, or complete. What if, in addition to the above four alternatives, several major companies would announce unexpected performance results? Should this be accounted for in the list of alternatives? No. Think why not, and see the answer in Note 3 at the end of the chapter.[3] These more complicated issues are introduced in the next section.

Gaussian or Normal Density. When observations of a continuous random variable x are affected by multiple random effects, in most cases the $\text{pdf}(x)$ has a specific functional shape that is called Gaussian. This fact is to a significant extent independent of the particular nature of the multiple random effects that contributes to the randomness of x. The exact formulation of this statement and its proof is a subject of the Central Limit Theorem (Cramer, 1946). Thus, Gaussian pdf plays a fundamental role in probability theory, and often is called a *normal* density. Most of this book is devoted to complicated cases, when a Gaussian shape is not appropriate for modeling pdfs. Nevertheless even in those cases, modifications or combinations of Gaussian functions will often be used. This is because when all deterministic sources of uncertainty are removed, the remaining random uncertainty is often normal, or Gaussian. Let us consider the basic properties of the Gaussian pdfs. For one scalar variable x, the Gaussian pdf, $G(x)$ is

$$G(x) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left[-0.5(x - M)^2/\sigma^2\right] \quad (1.2\text{-}4)$$

Here, M and σ are the parameters of the density: the mean and standard deviation. Also, σ^2 is called variance. Neither M nor σ is generally known and they have to be estimated from the data. The coefficient $(2\pi\sigma^2)^{-1/2}$ is defined so that $\int G(x)dx = 1$. We will also use notations $G(x|M, \sigma^2)$ for $G(x)$ to emphasize that this is the density, *given* the values of M and σ^2. A one-dimensional Gaussian density is illustrated in Fig. 1.2-2g.

Before we consider multidimensional densities, let us introduce basic vector and matrix notations. Vectors and matrices are used when a single event or object is characterized by several quantities. For example, a yesterday's Dow Jones closing is a single number x (called a scalar). But predicting its future values requires additional information, for example, a set of three values could be used (Dow Jones closing, interest rate, and gold price); such a set of three numbers is called a three-dimensional vector, denoted $\mathbf{x}$, or (x_1, x_2, x_3), or (x_i). Uncertainty of the predicted values of this vector can be characterized by using a covariance matrix, $\mathbf{C} = (C_{ij})$. Uncertainties of Dow Jones closing, interest rate, and gold price, when taken in isolation are characterized by C_{11}, C_{22}, and C_{33}, correspondingly; other elements of the covariance matrix characterize correlated uncertainties involving two variables.

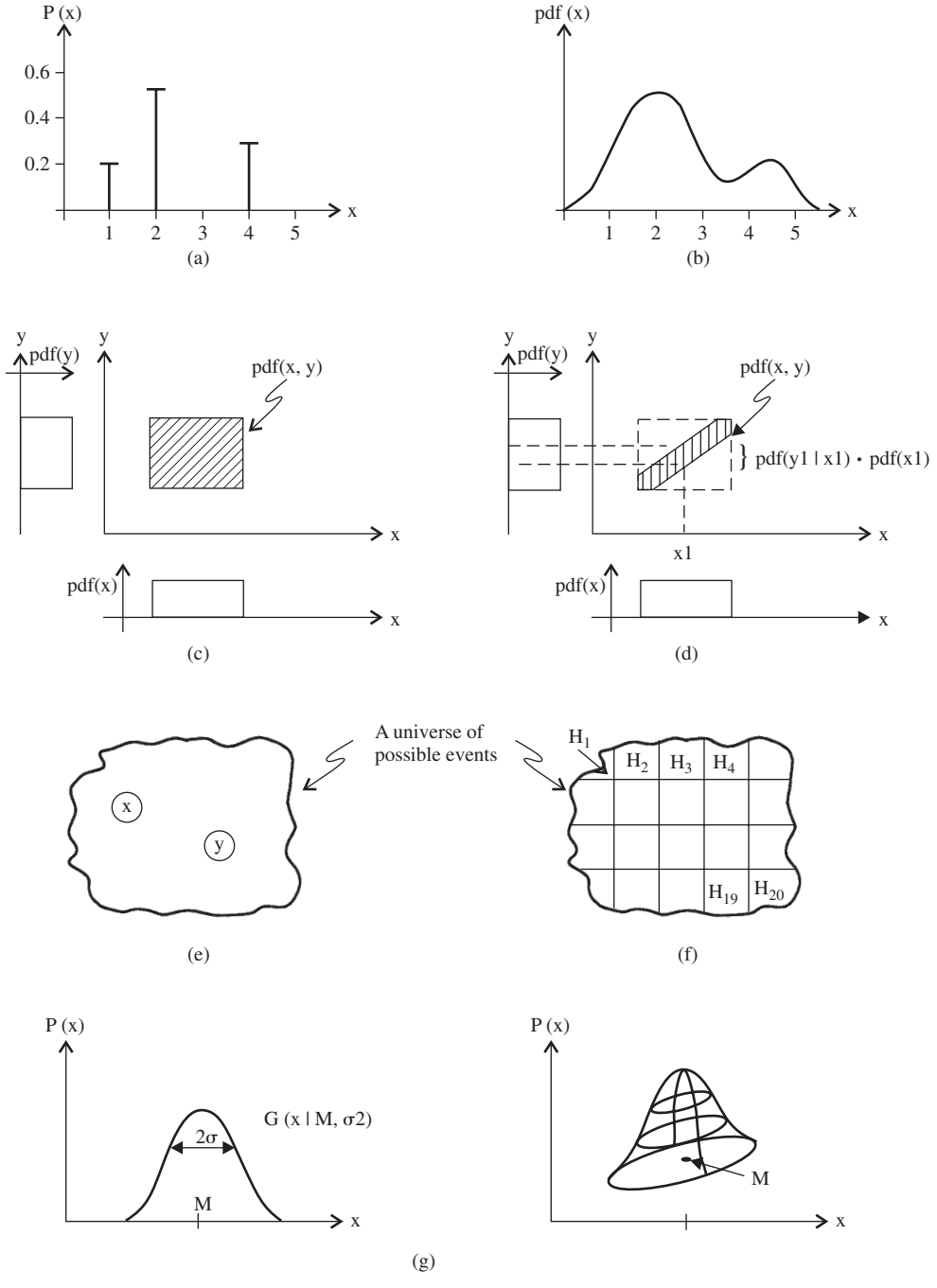


Figure 1.2-2 Illustration of basic notions of the probability theory. **(a)** probability of discrete events; **(b)** pdf of continuous events; **(c)** independent events; **(d)** dependent events and conditional probabilities; **(e)** alternative events; **(f)** a complete set of alternative events ($H_1 \cdots H_{20}$); **(g)** Gaussian distribution is a bell-shaped curve.

Vector and Matrix Notations. Multidimensional vector quantities $\mathbf{x}$ are one-dimensional arrays of scalar quantities, $\mathbf{x} = (x_1, x_2, \ldots, x_D)$; vectors and higher dimensional quantities are denoted in bold. Each x_d is called a component of the vector $\mathbf{x}$. The number of components, D, is called dimensionality. A matrix is a two-dimensional array,

$$\mathbf{C} = \begin{Bmatrix} C_{11} & C_{12} & \ldots & C_{1D} \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & C_{ij} & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ C_{D1} & C_{D2} & \ldots & C_{DD} \end{Bmatrix} \quad \text{or} \quad \mathbf{C} = (C_{ij}, i, j = 1, \ldots, D) \tag{1.2-5}$$

In two-dimensional case, components are often denoted x and y (instead of x_1 and x_2):

$$\mathbf{C} = \begin{Bmatrix} C_{xx} & C_{xy} \\ C_{xy} & C_{yy} \end{Bmatrix} \tag{1.2-6}$$

Definition. A vector is normally considered as a column; to denote a row vector, we use the transpose notation: $\mathbf{x}^{\mathrm{T}}$. For a matrix, transposition exchanges columns and rows: for $\mathbf{C} = (C_{ij})$, $\mathbf{C}^{\mathrm{T}} = (C_{ji})$.

Definition. Vector and matrix multiplication rule is "row by column":

$$\mathbf{Cx} = \left(\sum_j C_{ij} x_j \right), \quad \mathbf{x}^{\mathrm{T}}\mathbf{C} = \left(\sum_i x_i C_{ij} \right), \quad \mathbf{x}^{\mathrm{T}}\mathbf{y} = \left(\sum_i x_i y_i \right) \tag{1.2-7}$$

The above rule is called "inner product." Note that $\mathbf{xy}^{\mathrm{T}}$ is not an inner product,

$$\mathbf{xy}^{\mathrm{T}} = (x_i y_j) \tag{1.2-8}$$

It is called an outer product; it is a matrix.

Definition of Expected Value. The expected value of $\mathbf{x}$ is

$$E\{\mathbf{x}\} = \sum_{\mathbf{x}} \mathbf{x}\, P(\mathbf{x}) \quad \text{or} \quad \int \mathbf{x}\, \mathrm{pdf}(\mathbf{x})\, d\mathbf{x} \tag{1.2-8a}$$

Similarly, for any function $f(\mathbf{x})$

$$E\{f(\mathbf{x})\} = \sum_{\mathbf{x}} f(\mathbf{x}) P(\mathbf{x}) \quad \text{or} \quad \int f(\mathbf{x}) \mathrm{pdf}(\mathbf{x})\, d\mathbf{x} \tag{1.2-8b}$$

Definition of the Mean and Covariance. The expected value of $\mathbf{x}$ is called the mean of $\mathbf{x}$ (or mean value of $\mathbf{x}$, or mean $\mathbf{x}$), $\mathbf{M} = E\{\mathbf{x}\}$. Covariance of $\mathbf{x}$[Cov($\mathbf{x}$), or **Cov**, or $\mathbf{C}$] is

$$\mathrm{Cov}\{\mathbf{x}\} = E\left\{(\mathbf{x} - \mathbf{M})(\mathbf{x} - \mathbf{M})^{\mathrm{T}}\right\} = \int (\mathbf{x} - \mathbf{M})(\mathbf{x} - \mathbf{M})^{\mathrm{T}} \mathrm{pdf}(\mathbf{x})\, d\mathbf{x} \tag{1.2-8c}$$

Definition of the Average Value. Given observed data

$$\{\mathbf{x}_n, n = 1, \ldots N\} \tag{1.2-8d}$$

the average value of $\mathbf{x}$, $\mathbf{x}$-average or $\bar{\mathbf{x}}$, is

$$\bar{\mathbf{x}} = (1/N) \sum_n \mathbf{x}_n \tag{1.2-8e}$$

Similarly, the average value of $f(\mathbf{x})$, $\bar{f}(\mathbf{x})$ is

$$\bar{f}(\mathbf{x}) = (1/N) \sum_n f(\mathbf{x}_n) \tag{1.2-8f}$$

Definition of the Estimation and Estimator. Estimation is a process of finding approximate values for model parameters from the observed data. An estimation procedure is called an estimator.

To derive a good estimation procedure, we first must define what constitutes good estimation properties. This problem is considered in Chapter 4. Often, but not always, average values are good estimators for the corresponding expected values. We will sometimes use the same notations for parameters and their estimated values, say $\mathbf{M}$ for $\bar{\mathbf{x}}$.

Multidimensional Gaussian Density. For D-dimensional vector $\mathbf{x}$, Gaussian density is

$$G(\mathbf{x}|\mathbf{M}, \mathbf{C}) = (2\pi)^{-D/2} (\det \mathbf{C})^{-1/2} \exp\left(-0.5\, \mathbf{D}^{\mathrm{T}}\, \mathbf{C}^{-1} \mathbf{D}\right)$$
$$\mathbf{D} = \mathbf{x} - \mathbf{M}, \quad \mathbf{D}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{D} = \sum_{i;j} (x_i - M_i)\, C_{ij}^{-1} \left(x_j - M_j\right) \tag{1.2-9}$$

Here, det $\mathbf{C}$ is a determinant of the matrix $\mathbf{C}$, and $\mathbf{C}^{-1}$ is an inverse matrix of $\mathbf{C}$. In a two-dimensional case,

$$\det \mathbf{C} = C_{xx} C_{yy} - C_{xy}^2, \quad \mathbf{C}^{-1} = \begin{Bmatrix} C_{yy} & -C_{xy} \\ -C_{xy} & C_{xx} \end{Bmatrix} \Big/ \det \mathbf{C} \tag{1.2-10}$$

In higher dimensional cases, determinants and inverse matrixes are given by relatively complicated expressions or algorithms (for example, see Searle, 1982). We can use standard subroutines available in many software packages to compute these quantities. The shape of a D-dimensional Gaussian density can be illustrated by a D-dimensional contour of the bell curve computed at some constant pdf value, such a contour is an ellipsoid (see Fig 1.2-2h and Problem 1.2-1).

Elements of covariance matrixes C_{ij} are called covariances of x_i and x_j and the diagonal elements, C_{ii}, are called variances. Alternative notations are often used for the elements of covariance matrixes C_{ij}:

$$C_{ii} = \sigma_i^2, \quad C_{ij} = \sigma_i \sigma_j r_{ij} \tag{1.2-11}$$

where σ_i is called a standard deviation of x_i and r_{ij} is called a correlation coefficient between x_i and x_j. For example, if $x_1 =$ Dow Jones, and $x_3 =$ interest rate, a correlation coefficient value of $r_{13} = 0.3$ means that on average 30% of the Dow Jones variations go the same way as interest rate variations (and the other 70% go equally both ways).

Note that we used the same notations for $\mathbf{M}$ and $\mathbf{C}$ defined as expected values and as parameters of Gaussian densities. This requires justification (see Problem 1.2-2).

1.2.2 Classical Hypotheses Choice Paradigms and Definitions

In a classical hypothesis choice problem a decision has to be made based on available data. A decision consists in selecting one of several available hypotheses. Hypotheses represent the a priori knowledge, that is, the knowledge existing before the current piece of data is available. A decision is made a posteriori, that is, after the current piece of data became available. This area of probabilistic theory was developed by Bayes in 1763, and it is often called Bayesian decision theory. It was the first mathematical technique to combine a priori knowledge with data in the face of uncertainty, for making a posteriori predictions or decisions. Bayesian theory does not explain learning, still it represents one aspect of Aristotelian theory of Form: meeting between the a priori Forms (hypotheses) and matter (data).

In Bayesian theory each piece of data characterizing an object or a process is comprised of a set of measurements or vector $\mathbf{x} = (x_1, \ldots, x_D)$. The D-dimensional space $\{\mathbf{x}\}$ is called a decision or classification space. Hypotheses concerning objects, or in other words classes to which objects belong, are denoted as $\{H_1, \ldots H_K\}$ or simply, $k = 1, \ldots, K$. Each hypothesis or class is characterized by a priori probability, $P(H_k)$, and class-conditional probability density functions, $\text{pdf}(\mathbf{x}|H_k)$. Using the rule of conditional probabilities, Eq. (1.2-2), the joint density of (H_k and $\mathbf{x}$) is given by (see Problem 1.2-3)

$$\text{pdf}(H_k, \mathbf{x}) = P(H_k)\ \text{pdf}(\mathbf{x}|H_k) \tag{1.2-12}$$

These probabilities and pdfs are called a priori, because $P(H_k)$ and the functional expression for $\text{pdf}(\mathbf{x}|H_k)$ are known prior to data $\mathbf{x}$ being observed. [Probabilities $P(H_k)$ are called *priors* in classical statistics; but to emphasize that these quantities can be estimated from data we will usually call them *class rates* and use simplified notations $r_k = P(H_k)$.]

After data $\mathbf{x}$ is observed, what is the probability of each hypothesis? What are the *a posteriori* probabilities, $P(H_k|\mathbf{x})$? Using the rule of conditional probability, Eq. (1.2-2),

$$P(H_k|\mathbf{x}) = \text{pdf}(H_k, \mathbf{x})/\text{pdf}(\mathbf{x}) = P(H_k)\ \text{pdf}(\mathbf{x}|H_k)/\text{pdf}(\mathbf{x}) \tag{1.2-13}$$

Example. Consider all $\text{pdf}(\mathbf{x}|H_k)$ being of the same shape. Then, observations of $\mathbf{x}$ do not bring any additional classification information, and a posteriori probabilities remain equal to the a priori values, $P(H_k|\mathbf{x}) = P(H_k)$. (See also Problem 1.2-4 and the comment there.)

To complete the derivation of a posteriori probabilities from the measured data and a priori information, we need to express $\text{pdf}(\mathbf{x})$ in Eq. (1.2-13) in terms of a priori quantities. Since various hypothesis are alternatives, the total a priori $\text{pdf}(\mathbf{x})$ is a sum over the alternatives

$$\text{pdf}(\mathbf{x}) = P(H_1)\ \text{pdf}(\mathbf{x}|H_1) + P(H_2)\ \text{pdf}(\mathbf{x}|H_2) + \cdots + P(H_K)\ \text{pdf}(\mathbf{x}|H_K) \tag{1.2-14}$$

Combining Eqs. (1.2-13) and (1.2-14), we obtain the famous Bayes expression for a posteriori probabilities:

$$P(H_k|\mathbf{x}) = P(H_k)\ \mathrm{pdf}(\mathbf{x}|H_k)/\left[P(H_1)\ \mathrm{pdf}(\mathbf{x}|H_1) + \cdots + P(H_K)\ \mathrm{pdf}(\mathbf{x}|H_K)\right] \quad (1.2\text{-}15)$$

To repeat again, this Bayesian a posteriori probability was the first mathematical technique combining a priori knowledge with data for a posteriori predictions or decisions.

When choosing among hypotheses we often need to account for the cost and benefit of various decisions. If the cost/benefit of each decision in each case is known, and the cost $= -$benefit, this can be done as follows. Consider the benefit and cost matrixes

$$B(k|k') = -C(k|k') = \text{benefit of making the decision } k, \text{ when the actual class is } k' \quad (1.2\text{-}16)$$

[Presumably the benefit of the correct decision, $B(k|k)$ is the highest among $B(k|k')$ for $k \neq k'$; $C(k|k)$ is the lowest cost among $C(k|k')$.] A choice of hypothesis that maximizes the expected benefit is given by

$$\max_k \sum_{k'} B(k|k')P(H_{k'}|\mathbf{x}) \quad (1.2\text{-}17)$$

The fundamental significance of the Bayes theory is that any other hypothesis choice will result in a benefit less than (or the same as) that above. In the case of $B(k|k') = \delta_{kk'}(=$ 1 for $k = k'$, 0 otherwise), the best decision is given by the maximal a posteriori Bayes probability, $\max_k = P(H_k|\mathbf{x})$.

The choice between two hypothesis H_1 and H_2, in the case of $B(k|k') = \delta_{kk'}$, is given by $\max_k[B(k|k)P(H_k|\mathbf{x})]$. The decision rule can be written as

$$\begin{aligned} &\text{if } B(1|1)P(H_1|\mathbf{x}) > B(2|2)P(H_2|\mathbf{x}) \Rightarrow H_1 \\ &\text{if } B(1|1)P(H_1|\mathbf{x}) < B(2|2)P(H_2|\mathbf{x}) \Rightarrow H_2 \end{aligned} \quad (1.2\text{-}18)$$

This rule can be written as the likelihood ratio (LR) test:

$$\begin{aligned} &LR = \mathrm{pdf}(\mathbf{x}|H_1)/\mathrm{pdf}(\mathbf{x}|H_2) \\ &\text{threshold} = P(H_2)B(2|2)/[P(H_1)B(1|1)] \\ &\text{if } LR > \text{threshold} \Rightarrow H_1; \quad \text{if } LR < \text{threshold}, \Rightarrow H_2 \end{aligned} \quad (1.2\text{-}19)$$

Comment on Terminology. The word **likelihood** is used for a pdf, when it is considered for fixed data, as a function of hypotheses and their parameters. Although $\mathrm{pdf}(\mathbf{x}|H)$ was orginally used to describe probabilistic uncertainty of our knowledge about $\mathbf{x}$, given that it belongs to class H (is described by the model H), likelihood is used to characterize how well the observed data fit the model, or **how likely is our model, given the observed data**.

Example. Consider a more complicated example of the Bayes decision theory application. You are trading on a stock exchange. You have a model predicting the probabilities that the market will significantly move up or down within the next few days. That is, based on the information $\mathbf{x}$ available today, your model predicts a posteriori probabilities $P(\mathrm{up}|\mathbf{x}) = p$, $P(\mathrm{dn}|\mathbf{x}) = q$, and $P(\text{no move}|\mathbf{x}) = 1 - p - q$. (Note, this is a complete set of alternatives.) You would like to buy calls,[4] or puts, or do nothing. From your previous experience, you established the following sell rule: you close your position (that is you

sell all your calls or puts) in one of two cases: you gained 100% or you lost 30%. Given this rule, how should you formulate the optimal decisions? This means that you have to establish thresholds for the prediction probability, at which you will buy. This is done as follows. Your benefit/cost matrix is given by your sell rule

$$\mathbf{B} = \begin{Bmatrix} +1 & -0.3 & -0.3 \\ -0.3 & +1 & -0.3 \\ 0 & 0 & 0 \end{Bmatrix} \tag{1.2-20}$$

Here, rows correspond to hypothesis/actions (buy–call, buy–put, nothing) and columns correspond to actual outcomes (up, dn, no move). Say $B_{11} = 1$ corresponds to your buying calls and the market moves up, so you make a 100% gain (+1); if instead, the market moves down, you loose 30%, $B_{12} = -0.3$. Your probabilities corresponding to the three outcomes, k', form a vector $P(H_{k'}|\mathbf{x}) = (p, q, 1 - p - q)$. According to Eq. (1.2-17), the optimal decision rule is given by (see Problem 1.2-5)

$$\begin{aligned} &\max[(1.2p - 0.3), (1.2q - 0.3), 0], \text{ or equivalently, } \max[p, q, 0.23], \text{ so} \\ &\text{buy call if } p > q \text{ and } p > 0.23 \\ &\text{buy put if } q > p \text{ and } q > 0.23 \end{aligned} \tag{1.2-21}$$

1.2.3 Pattern Recognition

The importance of likelihoods or probabilities is due to the fact that the best decisions are based on these quantities. In a card game or stock market, the winner would be the one who uses for his decisions likelihoods or probabilities, if they are known. Similarly, when probabilities are known, the probabilistic framework provides for optimal formulation of any decision making. In the last example, any rule other than the Bayesian rule Eq. (1.2-21), would be less beneficial, if the probabilities p and q are correct.

However, applying probabilistic rules is usually complicated due to the fact that probabilities and pdfs are unknown and should be estimated from the data. The problem of applying the probabilistic rules is the central problem of the theory of probability; one could call it a "forward" problem. Estimating pdfs from data is the central problem of statistics; one could call it an "inverse" problem. In general, inverse problems are more complex than forward ones.

This complex problem of estimating pdfs is one of the central problems in mathematical methods of decision making. Often this problem gets replaced by other, simpler problems, because it is too complex. Another argument against estimating pdfs is that it requires a lot of data, and sometimes more direct goal optimization could be more efficient and could lead to a faster adaptivity. However, in the areas of hypothesis choice and recognition there are no general methods as powerful as the Bayesian one. An adaptive Bayesian method based on accurate estimation of pdfs, accounting for all the available information, if possible to implement, is usually the best approach. How to accomplish this, even in difficult cases, is discussed throughout this book.

The problem of hypothesis choice coupled with pdf estimation belongs to the area of pattern recognition. A large number of techniques have been developed for pdf estimation.

A mathematical analysis of the basic computational concepts is presented in Chapter 4. Here we introduce basic definitions and discuss some of the issues involved in this problem.

Definition. A *labeled* data set is a set of observations $\mathbf{x}_n, n = 1, \ldots, N$, each provided with an accurate label H_k. We denote a labeled data set $\{(\mathbf{x}_n, H_k)\}$, or, alternatively, we say that an observation $\mathbf{x}_n$ corresponds to a hypothesis H_k, or $n \in k$.

Definition. When a labeled data set is used for estimation of pdfs, learning (or training) is called *supervised*.

Definition. When a labeled data set is unavailable, the problem of learning is often formulated as clustering or grouping of observations $\{\mathbf{x}_n\}$ into "naturally occurring" clusters within the data set. Finding these groups or clusters is called *unsupervised* learning.

Sometimes, the concept of clustering is applied to *supervised* learning, when several clusters of data exist within each class. Thus, supervised and unsupervised learning become interwoven. A number of techniques have been developed to estimate pdfs, to cluster data, and to label clusters with class labels.

A widely used approach to supervised pdf estimation assumes a Gaussian shape of the class-conditioned pdfs, $\text{pdf}(\mathbf{x}|H_k) = G(\mathbf{x}|\mathbf{M}_k, \mathbf{C}_k)$. Then, a pdf estimation is reduced to the estimation of the pdf model parameters $\mathbf{M}_k$ and $\mathbf{C}_k$, which is accomplished using standard equations given in Chapter 4. This is called a parametric approach, because it is based on a parametric model of the pdf [Eqs. (1.2-4) or (1.2-9)].

A parametric approach utilizing a particular model is appropriate only if the model adequately represents the data. The Gaussian model is appropriate when there is a single deterministic phenomenon that determines the mean vector $\mathbf{M}_k$ of the density for each class and the deviations D_{xk} of the observations from this mean are of a random nature (Fig. 1.2-3a). In this case, the density is often Gaussian.[5] However, when there are two or more deterministic processes determining the means for each class, such as front and side views of an object, the density is likely to have two or more peaks in the classification space, deviating significantly from the Gaussian shape (Fig. 1.2-3b and c). To estimate pdfs of any shape, nonparametric methods of the pdf estimation were introduced. For example, in the Parzen method the pdf is estimated by placing kernel functions at each observed location $\mathbf{x}_n$ in the classification space. If a Gaussian kernel function is used, the pdf is estimated as follows:

$$\text{pdf}(\mathbf{x}|H_k) = \sum_{n \in k} (2\pi)^{-D/2} \, (\det \mathbf{C}_k)^{-1/2} \, \exp(-0.5 \, \mathbf{D}_{xn}^{\mathrm{T}} \, \mathbf{C}_k^{-1} \, \mathbf{D}_{xn})/N_k$$

$$\mathbf{D}_{xk} = \mathbf{x} - \mathbf{x}_n; \quad N_k = \sum_{n \in k} 1 \tag{1.2-22}$$

Here N_k is the number of observations from class k in the training data set. Nonparametric methods use many parameters (e.g., the location of kernel functions, $\mathbf{x}_n$),[6] whereas parametric methods use few parameters (e.g., the mean and covariance of the Gaussian density).

Nonparametric methods are efficient in classification spaces of low dimensionality, $D \sim< 6$. In high-dimensional spaces, learning requirements often grow fast as a function of the dimensionality. In fact, the number of observations in a training data set required for statistically accurate learning, N_k, often grows combinatorially or exponentially, $N_k \sim \exp(D)$. This contrasts with modest learning requirements of parametric methods: a general

rule for accurate estimation of the parameters of the Gaussian density [Eq. (1.2-13)] is $N_k \sim \max\left[10 \cdot D, D^2\right]$. Thus, parametric methods learn relatively fast. Fast learning or adaptation is at a premium in our ever-changing world; therefore, parametric methods are often used for practical real-world problems of medium dimensionality and reduction of dimensionality is desirable. But when dimensionality is high, training requirements of conventional parametric methods become prohibitive. More complicated parametric models accounting for more a priori information and using fewer parameters can adapt faster. These types of models and their estimation techniques are considered throughout the book.

1.2.4 A Priori Information and Adaptation

Relative roles of a priori information and adaptation have been debated among philosophers during the past 2300 years. This issue has continued to be among the most controversial ones during the past 50 years among researchers of intelligence. It acquired various names such as internal representations, symbolic or discrete methods (a priori) vs. neural, connectionist, or continuous (adaptive). In classical statistical pattern recognition, adaptivity is accomplished by pdf estimation; in the case of the parametric models, such as Gaussians, adaptivity is due to adaptive parameters of pdfs, which are estimated using available data. Learning is called supervised when pdf are estimated using a labeled data set: each object is assigned to a known class. A labeled data set is prepared using a priori knowledge, so that there is no real-time adaptation. In unsupervised learning, adaptation occurs in real time, as data become available. A priori information includes a training data set, knowledge of the parametric shape of pdf or kernel functions, and knowledge of classification features. Features are functions of observables that are designed or selected in such a way that the important classification information is preserved, while the dimensionality of the problem is reduced. The components of the data vector $\mathbf{x} = (x_1, \ldots, x_D)$ can be defined as raw sensory data or as features computed during preprocessing.

The importance of feature selection follows from the above discussion of training requirements. Consider typical problem dimensionalities. For example, a stock market prediction may depend on the market values over the previous 30 days, also on interest rates over the past 30 days, on closings of several international markets, and several other indicators computed for 30 days each; so, the desire to account for more information quickly leads to hundreds of dimensions. Similarly, a radar target signature may contain hundreds of time-point measurements, and at every time point, amplitudes and phases at multiple frequencies might be measured; even more pixels are usually contained in imaging sensory data. Thus, even a modest requirement of D^2 training signatures per class is often prohibitive.

Feature design and selection are based on careful analysis of a priori phenomenological or scientific information about the data as well as on the analysis of statistical properties of data and features. A large amount of technical literature is devoted to feature selection, which remains to a significant extent application specific. Considerations of invariances with respect to translation, rotation, etc. are often used to design features for image data. Discussion of general feature evaluation methods can be found in Fukunaga (1972, 1991); these methods are useful for subselection of features out of an initial set of features. It should be pointed out that from a statistical theoretical point of view, there always exists a "superfeature," a single feature containing all the classification information. A likelihood

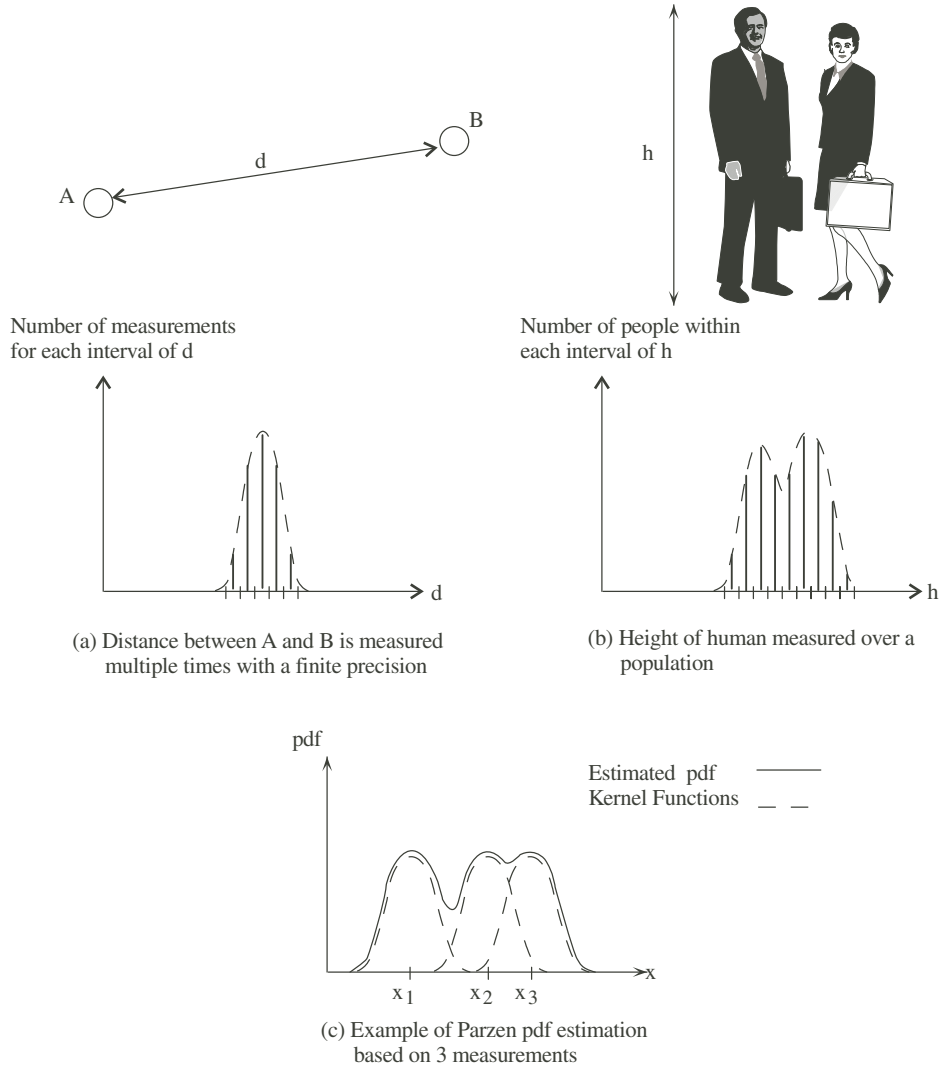


Figure 1.2-3 A single class density can be modeled as Gaussian, if there is a single deterministic phenomenon that determines the class mean and the deviations are random (a). When there are two deterministic processes within each class, such as male and female heights (b), front and side views of an object, or diverse market forces (c) the density is likely to have two or more peaks in the classification space, deviating significantly from the Gaussian shape.

ratio, or the Bayes classifier, or a sufficient statistics[7] of the sensory data are such features, and, sometimes, useful features can be constructed by approximating sufficient statistics. However, general automatic methods of designing features have not been found in classical statistical pattern recognition. As already mentioned, statistical pattern recognition models

the adaptive nature of the intellect; however, it has not succeeded in modeling the a priori nature of the intellect.

An entirely different approach to utilizing a priori information for decision making and pattern recognition is utilized in rule-based or expert systems. In the Plato–Minsky method, classification decisions are specified as a series of "if–then" rules. The main advantage of this method is that it explicitly incorporates detailed, high-level a priori knowledge into the decision making. The main drawback of this method is difficulty of combining rule systems with adaptive learning as discussed in the previous section. Extensions of the rule-based concept to two-dimensional (2-D) and three-dimensional (3-D) sensory data led to another mathematical concept of utilizing the a priori information, model-based vision. Model-based approaches in machine vision utilize detailed a priori information on objects' shape for image recognition and understanding. Models used in machine vision typically are complicated geometric 3-D models. To find and identify objects in a scene, first, one has to extract a subset of the image that corresponds to a single object; this is called grouping, selection, or segmentation. Second, one has to match the extracted data to a model from a database of models; this is called matching.

If the position, size, and orientation of an object in the image are known with little uncertainty, model-based vision works well. But if there is uncertainty, the solution becomes combinatorially complex because it is necessary to consider all or many of the possible combinations of all the parameters involved: subsets of imagery data, object models, object position, and orientation. For example, consider a medium complexity problem of recognizing one of 100 possible objects that should be identified in an image of 1000×1000 pixels. If we know the size and orientation of every object, the only unknowns are the position and type of the object. One needs to evaluate no more different positions than there are pixels, $\sim 10^6$. There are 100 possible objects that should be matched to the image in every position. The number of all possible matches to be evaluated in this example is $\sim 10^6 \times 100 \sim 10^8$, which is a very large number. And even this number is limited by our assumptions that we know the size of the object in the image, that there is no obscurations or other uncertainties due to illumination, deviation of objects from their models, etc. When many objects obscure each other, this leads to uncertainty in segmentation: which pixel belongs to which object. It is necessary then to consider many possible different segmentations of every little part of an image; if two objects may occupy a 100×100 pixel subimage, using brute force may lead us to consider a significant part of all possible segmentations, on the order of 100^{100}. This number is comparable to the number of all interactions of all elementary particles in the entire history of the universe and certainly that many matches cannot be implemented in any computer. In fact, the uncertainties require more options in the matching or more detailed models, potentially leading to a combinatorial explosion at *every* one of the model-based vision steps discussed above: "the key issues . . . the inherent uncertainty of data measurements" and "combinatorial explosion inherent in the problem" (Grimson and Huttenlocher, 1991).

Nevertheless, the concept of an internal model is among the most important in the mathematics of intelligence, because it combines the apriority of a model with the adaptivity of model parameters. Developing mathematical methods that utilize internal models, while avoiding combinatorial explosion, is among the main thrusts of this book. As an introduction to this development, the next subsection considers a mathematical formulation of the internal model and analyzes the emerging difficulties.

1.2.5 Mathematical Formulation of Model-Based Recognition

The following discussion refers to pixels and images, but it is equally applicable to virtually any type of data. In case of images, we use geometric models and features such as edges or corners, and in case of stock market, we use models and features such as bottoms and tops.

A mathematical formulation of recognition based on internal models consists of two steps: first, the a priori step of model development and second, the adaptive step of recognition of objects in real-time data, while adapting unknown parameters of the models. Models of data should be developed for every class k of objects. We denote models for class k as $\mathbf{M}_k$ and the nth data vector as $\mathbf{x}_n$. Models represent an expected, deterministic aspect of the data. In the case of perfect models and no uncertainties, when the data $\mathbf{x}_n$ originate from an object of class k, the data perfectly matches the model,

$$\mathbf{x}_n = \mathbf{M}_k \tag{1.2-23}$$

When there are deterministic uncertainties, the models are functions of unknown model parameters, $\mathbf{S}_k$ (such as orientation), $\mathbf{M}_k = \mathbf{M}_k(\mathbf{S}_k)$. And the above expression is understood as an equation for model parameters: there are values of parameters $\mathbf{S}_k$ such that Eq. (1.2-23) holds. For imaging data, every data vector $\mathbf{x}_n$ is a subset of image pixels, or features extracted from a subset of the image. The model is a prediction of this vector (of pixels or features) and it should account for the properties of objects and the sensory system. In reality, a perfect match as in Eq. (1.2-23) cannot be attained because there are multiple sources of deviations between the model and the data. If all deterministic aspects of the problem are accounted for in the model $\mathbf{M}_k(\mathbf{S}_k)$, the deviations can be treated statistically; that is, for some values of the parameters $\mathbf{S}_k$, the model is a class-conditional statistical expectation of the data vector

$$\mathbf{M}_k(\mathbf{S}_k) = E\{\mathbf{x}_n|k\} \tag{1.2-24}$$

The expectation here is conditioned on k, rather than on the class-k hypothesis H_k, to emphasize that it is taken with respect to the true $\text{pdf}(\mathbf{x}_n|k)$, which is not known. Let us emphasize again that the above equation is not a definition of the model, but the goal that the model should satisfy.

If the deterministic model $\mathbf{M}_k(\mathbf{S}_k)$ is accurate in the sense of Eq. (1.2-24), and deviations of data from the model are caused by multiple random effects, then according to the Central Limit theorem, the pdf can be approximated by a Gaussian density, conditioned on class and model parameter values,[8]

$$\text{pdf}(\mathbf{x}_n|H_k) = G\,[\mathbf{x}_n|\mathbf{M}_k(\mathbf{S}_k), \mathbf{C}_k(\mathbf{S}_k)] \tag{1.2-25}$$

Here, G is Gaussian density, Eq. (1.2-9), with the mean given by the deterministic model and the covariance $\mathbf{C}_k(\mathbf{S}_k)$, which should either be modeled deterministically or estimated from the data. This formulation, combining deterministic models and statistical uncertainties, relates the model-based recognition problem to the classical problem of hypothesis choice considered in Section 1.2.2.

The next step, adaptive object recognition, consists in finding the class k and the values of parameters $\mathbf{S}_k$ that result in the best match in some sense consistent with Eqs. (1.2-23) or

(1.2-24). This problem can be formulated mathematically as follows. Consider all possible associations (also called segmentations or partitions) of data among all classes or hypothesis. Association mathematically can be described as a partition Ξ of all the data (pixels or samples) into subsets ξ_n corresponding to particular objects,

$$\Xi = \{\xi_1, \ldots, \xi_N\} \tag{1.2-26}$$

Pixels from a subset ξ_n are used to form a data vector $\mathbf{x}_n$ to which the model $\mathbf{M}_k$ is then matched. Here, $n = 1, \ldots, N$ numbers the objects, $k = 1, \ldots K$ numbers the classes of objects, and the partition Ξ establishes the correspondence between them, $k(n)$. An adaptive matching of the models to data involves estimation of the unknown parameters of the models, $\mathbf{S}_k$, for every subimage ξ_n. A standard preferable approach to parameter estimation is to maximize the likelihood, L, or overall pdf of all pixels in the image. Theoretical advantages of likelihood maximization are well known and will be discussed in Chapter 4. To obtain an expression for the likelihood, we need to examine the probabilistic interpretation of the segmentation (1.2-26). The global hypothesis corresponding to this partition states that the following events occur *concurrently:* object $k(1)$ is observed in pixel subset $\xi_1, \ldots,$ object $k(n)$ is observed in pixel subset ξ_n, etc. Models account for any deterministic variability, whereas deviations from the models are random and statistically independent for different subsets. For statistically independent events, the joint pdf is a product of individual pdfs. Thus, the likelihood conditioned on a particular segmentation, $L(\Xi)$, is a product of conditional pdfs, pdf $\left(\mathbf{x}_n | H_{k(n)}\right)$,

$$L(\Xi) = \prod_n \text{ pdf } \left(\mathbf{x}_n | H_{k(n)}\right) \tag{1.2-27}$$

It is worth emphasizing that although Eq. (1.2-27) is a product of individual subset pdfs, it does not require an assumption of statistical independence of $\mathbf{x}_n$ and $\mathbf{x}_{n'}$: they can always be made statistically independent with the appropriate selection of models, which may account for intersegment dependencies through the models.[9] This will be further discussed in Chapters 4 through 7.

The maximum likelihood estimate of the set of parameters $\{\mathbf{S}_k\}$ is obtained by maximizing expression (1.2-27) over the parameters, which we will denote as

$$\{\mathbf{S}_k\} = \arg \max \ L(\Xi) \tag{1.2-28}$$

According to (1.2-27), this problem factors into maximizing pdf of each individual subimage over the parameters of each individual model

$$\mathbf{S}_{k(n)} = \arg \max \text{ pdf}(\mathbf{x}_n | H_{k(n)}) \tag{1.2-29}$$

This greatly simplifies the problem of parameter estimation, which still could be quite substantial for complex models. After the best set of model parameters is obtained for every partition, the likelihood for every partition is computed and parameter values corresponding to the maximum likelihood partition are selected.

The above formulation of the model-based pattern recognition problem is fairly broad. It addresses the top level of the problem, while omitting details important for specific

application areas and for specific approaches to controlling the combinatorial explosion. This general formulation will be referred to as the Multiple Hypothesis Testing (MHT) algorithm, designating the fact that multiple partitions of data among hypotheses have to be tested to find the maximum likelihood partition and parameters. Most of existing approaches can be formulated within the MHT framework. The model-based pattern recognition is a step toward mathematical description of the Aristotelian theory of mind. The functional shapes of models, $\mathbf{M}_k(\mathbf{S}_k)$, are specified a priori and correspond to the Aristotelian Forms. The adaptation or learning is achieved by estimating model parameters, $\mathbf{S}_k$, from the data; this process mathematically describes the meeting of the Form and matter. However, the need to consider all or many of the partitions or groupings (1.2-26) is the main source of the intrinsic combinatorial complexity of the model-based approach. Overcoming this complexity is considered in Chapter 4.

1.2.6 Conundrum of Combinatorial Complexity

Combinatorial complexity seems to be an omnipresent fixture in every computational concept. Attempts to develop self-learning, adaptive algorithms utilizing no a priori knowledge faced combinatorial complexity of the amount of required training data. Rule-based approaches, which were proposed to overcome this difficulty by using the concept of apriority, faced combinatorial complexity of rule systems. And attempts to combine apriority and adaptivity using model-based approaches faced computational combinatorial complexity. The conundrum of combinatorial complexity will be further analyzed in Chapter 2, where it will be related to the general properties of Aristotelian logic. And, in Chapter 10, it will be discussed in relationship to the Gödel theorems. In Chapter 4, we develop a computational concept of modeling field theory that resolves this conundrum using adaptive fuzzy logic.

1.3 Prediction, Tracking, and Dynamic Models

Here we introduce basic definitions, summarize classical approaches, and discuss relationships among prediction, tracking, and pattern recognition. Then we consider the general problem of prediction in complicated cases when dependencies among variables are nonlinear and data, in addition to signals of interest, contain noise and clutter (distracting signals). This section serves as an introduction to Chapter 7.

In the beginning of every mythology, theology, or cosmogony, there is a concept of the original chaos. An emergence of ordered cosmos is equated with the divine act of creation, which psychologically is equivalent to an emergence of consciousness. An ability to order, to predict is considered fundamental to consciousness. Ancient Greeks' passion for mathematics was related to the psychological need to counter chaos, and the same psychological need moves many of us today. Ancient mathematical concepts of arithmetic and geometry countered chaos by establishing deterministic relationships among mathematical objects. This deterministic mathematics was used for prediction in astronomy, where deterministic predictions work very well. Beginning with the sixteenth century, more sophisticated mathematical methods emerged, rationalizing the chaos itself. Prediction of outcomes in card games and gambling stimulated the development of probability theory,

which was originally developed for predicting random combinations of a few basic simple events. Next, prediction methods were developed for continuous variables. These eventually led to mathematical techniques combining the deterministic and probabilistic aspects of nature.

The most basic and widely used prediction method that combines probabilistic and deterministic aspects is linear regression. It is used to establish linear relationships among variables when such a relationship is of a stochastic nature and does not hold precisely for every set of measurements, but is observed probabilistically over a large set of observations. (We use the words *measurements* and *observations* interchangeably.) Regression can be used to establish relationships between the past and future values of the same variable. A set of measurements obtained over regular (or irregular) time intervals is called a time series, and a linear regression model applied to time series is called autoregression. Autoregression modeling methods were developed in the first half of this century. Further development of prediction methods was affected by two events: the development of stochastic process theory, and the need to solve the problem of target tracking that came up during World War II, when radar was used to track aircraft. Today, tracking applications are numerous in both military and civilian areas of surveillance, navigation, guidance, air traffic control, and robotics. To track targets or objects it is necessary to be able to detect the presence of a target and to predict where the target will appear in the next moment. These detection and prediction aspects of tracking are considered in this book. Tracking problems are characterized by complex models: deterministic tracking models are often nonlinear, and probabilistic tracking models often account for multiple sources of signal: multiple targets of interest as well as multiple sources of noise and cutter. Sophisticated tracking methods are utilized for prediction, in particular, of stock markets.

Three hundred years ago, prediction of outcomes in card games and dice gave an impetus to the development of the theory of probability. Today, new mathematical methods of prediction are being developed for predicting stock markets. Identification of investment opportunities is crucial for the efficient economy and enriches those who can predict better than average state-of-the-art techniques.

In this section we introduce the mathematics of linear regression in the most simple case of two variables. Then we briefly consider autoregressive modeling and tracking, emphasizing similarities and differences among regression, tracking, and pattern recognition.

1.3.1 Linear Regression

Consider an estimation of unknown values of variables $\mathbf{y}$ from known values of variables $\mathbf{x}$. For example, $\mathbf{x}$ may include time, known past values of $\mathbf{y}$ (autoregression), or other variables useful for predicting $\mathbf{y}$. Classical linear regression estimates a linear relationship between $\mathbf{x}$ and $\mathbf{y}$ from available past observations of pairs $(\mathbf{x}, \mathbf{y})$. Here we consider a simple case of one (scalar) known variable x and one scalar unknown variable y; a general case of vector-valued $\mathbf{x}$ and $\mathbf{y}$ is considered in Chapter 7. Linear relationship between x and y is given by

$$y = ax + b \tag{1.3-1}$$

where a and b are the parameters of the regression model that should be estimated from the available data

$$\{(x_n, y_n), \quad n = 1, \ldots, N\} \tag{1.3-2}$$

In the real world, Eq. (1.3-1) cannot hold exactly, because of measurement errors and other random effects. Thus, we have to account for the error,

$$\varepsilon_n = y_n - ax_n - b \tag{1.3-3}$$

This error can often be considered a random variable, with the same probability density $\text{pdf}(\varepsilon_n) = \text{pdf}(y_n - ax_n - b)$ for every n. If the data (x_n, y_n) for every n are affected by statistically independent random effects, then the joint pdf or likelihood of the data $\{(x_n, y_n)\}$ is given by a product of individual $\text{pdf}(y_n - ax_n - b)$ over n. It is more convenient to consider the logarithm of the likelihood (log likelihood), $LL = \ln\, L$,

$$L = \prod_n \text{pdf}(y_n - ax_n - b); \quad LL = \ln\, L = \sum_n \ln\, \text{pdf}(\varepsilon_n),$$
$$\varepsilon_n = y_n - ax_n - b \tag{1.3-4}$$

This likelihood combines the deterministic model (1.3-1) with a probabilistic model given by $\text{pdf}(\varepsilon)$. Parameters a and b of the regression model can be obtained by maximizing LL over a and b. Classical linear regression considers $\text{pdf}(\varepsilon)$ to be a Gaussian density, (1.1-3), so the log likelihood is given by

$$\begin{aligned} LL &= \sum_n \ln\left[\left(2\pi\sigma^2\right)^{-1/2} \exp\left(-0.5\, \varepsilon_n^2/\sigma^2\right)\right] \\ &= \sum_n \left\{-0.5\, \ln\left(2\pi\sigma^2\right) - 0.5\, \varepsilon_n^2/\sigma^2\right\} \end{aligned} \tag{1.3-5}$$

Here, the first item in the parentheses does not depend on the regression parameters a and b. Therefore, for the maximum likelihood estimation of the regression parameters, it is sufficient to consider just the second item,

$$\max_{a,b} \left\{\sum_n \left(-0.5\, \varepsilon_n^2/\sigma^2\right)\right\} \tag{1.3-6}$$

or equivalently,

$$\min_{a,b} \left\{\sum_n \varepsilon_n^2\right\}, \quad \varepsilon_n = y_n - ax_n - b \tag{1.3-7}$$

Because of the sum of squares in this expression, this technique is often called the least mean square method. To find this minimum, the derivatives with respect to a and b are equated to 0:

$$\begin{aligned} d/da \left\{\sum_n \varepsilon_n^2\right\} &= d/da \left\{\sum_n (y_n - ax_n - b)^2\right\} = -2 \sum_n (y_n - ax_n - b)\, x_n = 0 \\ d/db \left\{\sum_n \varepsilon_n^2\right\} &= d/db \left\{\sum_n (y_n - ax_n - b)^2\right\} = -2 \sum_n (y_n - ax_n - b) \qquad = 0 \end{aligned} \tag{1.3-8}$$

This is a linear system of equations for a and b and it can be solved in a straightforward manner. Let us introduce the following notations: x-average, $\bar{x}$, and y-average, $\bar{y}$,

$$\bar{x} = (1/N)\sum_n x_n, \quad \bar{y} = (1/N)\sum_n y_n \tag{1.3-9}$$

variance of x, C_{xx}, and covariance of x and y, C_{xy}

$$C_{xx} = (1/N)\sum_n (x_n - \bar{x})^2, \quad C_{xy} = (1/N)\sum_n (x_n - \bar{x})(y_n - \bar{y}) \tag{1.3-10}$$

Note, that we are using here the same notations as we used for the covariance matrix of the Gaussian densities in Eqs. (1.2-9) and (1.2-10). These quantities are closely related as explained in Section 1.2.1 and in the next section. With simple manipulations (see Problem 1.3-1), Eqs. (1.3-8) can be rewritten as

$$\left.\begin{aligned} C_{xy} - aC_{xx} &= 0 \\ \bar{y} - a\bar{x} - b &= 0 \end{aligned}\right\} \tag{1.3-11, 1.3-12}$$

From here, we obtain parameters of the linear regression model:

$$a = C_{xy}/C_{xx}, \quad b = \bar{y} - \left(C_{xy}/C_{xx}\right)\bar{x} \tag{1.3-13}$$

Substituting these expressions into Eq. (1.3-1), we obtain the linear regression model:

$$y(x) = x\, C_{xy}/C_{xx} + \bar{y} - \left(C_{xy}/C_{xx}\right)\bar{x} = \bar{y} + (x - \bar{x})C_{xy}/C_{xx} \tag{1.3-14}$$

Or, by using notations of Eq. (1.2-11),

$$y(x) = \bar{y} + (x - \bar{x})\, r_{xy}\sigma_y/\sigma_x \tag{1.3-15}$$

This expression can be interpreted as follows: $y(x)$ equals y-average plus the rescaled deviation of x from x-average; the rescaling factor, $r_{xy}\sigma_y/\sigma_x$, accounts for the correlation between x and y and for the difference in scales between x and y, as measured by the standard deviations of x and y.

1.3.2 Regression as an Expectation

The previous section introduced regression as an estimation of the parameters of a deterministic model. Another fundamentally important point of view, which connects regression and pattern recognition, is to consider regression as an expectation of a y value, given a value of x. The mathematical expectation $E\{y|x\}$ (read: "expected value of y, given x") is defined as

$$E\{y|x\} \equiv \int y\ \mathrm{pdf}(y|x)dy \tag{1.3-16}$$

Here, $\mathrm{pdf}(y|x)$ is a conditional density of y, given x. This conditional density can be

defined through the joint density of x and y, $\text{pdf}(x, y)$, and unconditional density of x, $\text{pdf}(x)$, according to the rule of conditional probabilities,

$$\text{pdf}(x, y) = \text{pdf}(y|x)\text{pdf}(x), \quad \text{or } \text{pdf}(y|x) = \text{pdf}(x, y)/\text{pdf}(x) \tag{1.3-17}$$

The meaning of this expression is very simple: $\text{pdf}(y|x)$ is proportional to $\text{pdf}(x, y)$, and the denominator in the above expression assures the proper pdf normalization: $\int \text{pdf}(y|x)\, dy = 1$; this can be seen from $\text{pdf}(x) = \int \text{pdf}(x, y)\, dy$.

Consider joint $\text{pdf}(x, y)$ to be Gaussian, (1.1-3). Substituting Gaussian densities for $\text{pdf}(x, y)$ and $\text{pdf}(x)$ in (1.3-16 and 1.3-17) (see Problems 1.3-4 and 1.3-5), we obtain the same expression as in the previous section, (1.3-14):

$$E\{y|x\} = \bar{y} + (x - \bar{x})C_{xy}/C_{xx} \tag{1.3-18}$$

Note that in Eq. (1.3-18) averages and covariances are the true theoretical values of the density parameters, whereas in the previous section we used their values estimated from the data. Obviously, in practice the "true theoretical values" of parameters are not known and, using the regression Eq. (1.3-18) for prediction requires estimated parameters of pdfs.[10] This need to estimate pdf unifies regression and pattern recognition.

Summarizing this section, we can say that Gaussian densities imply linear relationships between variables. In Chapter 7 we consider more complicated densities leading to nonlinear relationships.

1.3.3 Autoregression

Consider a time series $\{x_t, t = 1, \ldots, N\}$. For simplicity we use integer time values, but use the index t instead of our usual n to emphasize the nature of the data as a time series. And let us consider the problem of predicting x_{t+1} from x_t. We will use liner regression:

$$x_{t+1} = ax_t + b \tag{1.3-19}$$

This is equivalent to assuming that x_{t+1} is affected only by x_t, and that the joint density is Gaussian. This model is called autoregressive, and a is called the autoregressive coefficient. The solution for the coefficients a, b is given by Eq. (1.3-15)

$$x_{t+1}(x_t) = r\,(\sigma_{t+1}/\sigma_t)\;(x_t - \bar{x}_t) + \bar{x}_{t+1} \tag{1.3-20}$$

Note that here $\sigma_x = \sigma_t$ and $\sigma_y = \sigma_{t+1}$, and we denoted $r_{t,t+1}$ as r. In the previous two sections we assumed that we have multiple observations of the pairs of (x, y) from previous experience. This allowed us to estimate parameters of statistical models: means, standard deviations, and correlations (parameters of statistical models are also called statistics). Without this assumption, regression will be useless. For a time series, we never have the same x and y, as both of them are changing with time. So how can we obtain statistics, $\bar{x}_t, \bar{x}_{t+1}, r, \sigma_t, \sigma_{t+1}$, which are needed to use Eq. (1.3-20) for prediction? For a time series, it is often the case that although x_t are changing significantly and randomly from t to $t + 1$, statistics of this random process change slowly. An idealization of this observation is formulated as

stationarity assumption: assume that statistics do not change in time (1.3-21)

So that $\bar{x}$, r, and σ remain constant, $\bar{x}_t = \bar{x}_{t+1}$ and $\sigma_t = \sigma_{t+1}$. Then, the autoregression Eq. (1.3-20) simplifies,

$$x_{t+1}(x_t) = r(x_t - \bar{x}) + \bar{x}, \quad \text{or } x'_{t+1}(x_t) = rx'_t \tag{1.3-22}$$

where the normalized zero-mean variables are introduced: $x'_t = x_t - \bar{x}$, $x'_{t+1} = x_{t+1} - \bar{x}$. This model is more specifically designated as the first-order autoregressive model (it looks just one step back), or a random walk model. It predicts that tomorrow's x value will be closer to the average than today's value. It can be used for prediction, for example, of the stock market (see Problem 1.3-8). Statistics, $\bar{x}$, r, and σ, should be estimated from the past data, going as far back in the past as one believes in the stationarity assumption. The stationarity assumption can be verified by comparing statistics over various time intervals (see Problem 1.3-9).

For stock market predictions, the basic assumptions of the model are too simplistic. There are two assumptions: stationarity and the first-order single-variable autoregressive model. Concerning stationarity, it is well known that the average stock price is steadily rising, so this assumption is wrong. Instead of stock prices let us consider their changes, $x'_t = x_t - x_{t-1}$. The average value of this variable is very close to 0. Therefore, it is better to use x' than the original x, as x' is closer to the assumption of stationarity. The first-order autoregressive model for x' contains information about yesterday's, in addition to today's price: this model predicts that the change in price tomorrow will be similar to today's change if $r > 0$, or the opposite if $r < 0$ (see Problem 1.3-10).

Still, x' usually is not stationary because the autoregression (correlation) coefficient and standard deviation may change with time. In some market conditions, it is positive: the next day change tends to be in the same direction as today's change. In other market conditions, change tends to be contrarian ($r < 0$). Therefore, the correlation coefficient has to be estimated from the minimum amount of the past data to be as current as possible (see Problem 1.3-11). But reducing the past interval too much will result in inaccurate (statistically unreliable) estimation. Also, predicting types of markets ($r > 0$ or $r < 0$) may require other variables (say interest rates); thus, one has to consider predictive models with multiple variables. Using multiple variables increases dimensionality and requires even more data for estimation of statistics. This is an inherent difficulty of prediction problems.

The prediction methods just described are practically useful and mathematically optimal in very many cases. Could they be used for stock market prediction? And if not, why are they not good enough? The answer is no, and the reason is very simple: these methods are widely known and constantly used by very many investors. So whatever could be predicted with these methods becomes immediately predicted, money from the entire financial community is redistributed more efficiently, and the potential for above-average gains for individual investors is reduced. When the previously described methods of prediction, based on stationary linear models, were introduced in the financial markets, they first resulted in above-average gains, and second, in increased market efficiency. Increased market efficiency eliminated formerly predictable stationary linear effects, so whatever could be predictable now has to be nonstationary, nonlinear, and, in other ways, more

complicated than the state of the art. Thus it is important to identify principled limitations of existing methods.

Among limitations of the autoregression model, even with many variables and of a high order, is a restricted nature of its adaptivity. Autoregression models, as well as regression models in general, assume that there is a single deterministic process determining the mean of the future price and that other effects are random deviations from the mean. The assumption of the Gaussian density of the deviations further restricts adaptivity to linear combinations of input variables. But the stock market is not linear, and it is affected by a number of dynamic processes or forces acting concurrently. The next section introduces classical prediction methods developed in the area of target tracking. Target tracking combines pattern recognition and prediction, and provides a springboard to developing prediction methods that can adapt to new forms of nonlinearities and account for multiple concurrent processes.

1.3.4 Tracking

Powerful techniques combining prediction with pattern recognition have been developed in the field of target tracking that concern detection of moving objects and estimation of their trajectories in sensory data. Sensor measurements typically contain data that originate from multiple sources: objects of interest called targets, objects of no interest called background or clutter, and sensor noise. In simple cases, when there is a single target and no noise or clutter, the data can be considered as a time series of coordinate measurements. Tracking is similar to prediction problems considered previously: it consists in estimating parameters of motion models (Fig. 1.3-1a). Complicated tracking problems involving multiple targets, noise, and clutter have been traditionally approached by subdividing them into several simpler steps, called functions: detection, association, and track estimation. Detection refers to the process of determining samples or pixels of data containing target signals while rejecting clutter and noise. Association refers to grouping of data from multiple frames into subsets corresponding to a single object. And track estimation refers to estimating parameters of the model of target motion (position, velocity, etc.) (see Fig. 1.3-1b).

Detection and association functions correspond to the segmentation step in computational vision. Historically, mathematical methods of tracking and computer vision evolved

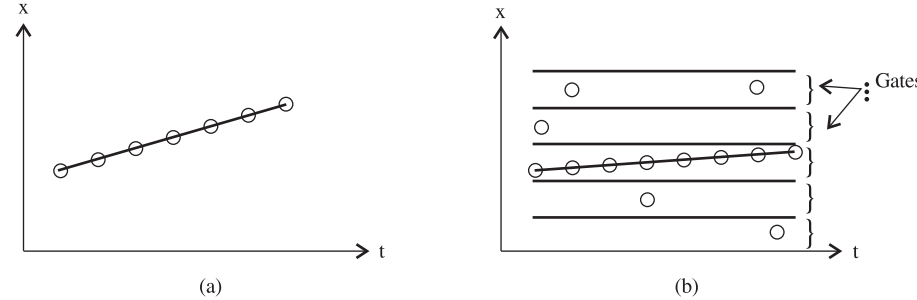


Figure 1.3-1 Tracking is easy for a single target without noise or clutter **(a)**. Association is required when multiple targets or clutter signals are present **(b)**; association is performed using gates.

along different routes, but today they are beginning to merge. Tracking begins to be concerned with imagery data, and vision is concerned with image flow and motion. In computational vision, a segmentation is naturally the first step to be performed, whereas object models are complicated and have been incorporated into recognition only recently. In simple tracking cases, track-model parameters can be estimated without association, and model-based tracking algorithms were quickly developed. Combining tracking with association was developed only much later. Today, powerful tracking techniques are being applied in the area of prediction, in particular, stock market prediction, where identification and estimation of multiple concurrently acting forces mathematically resemble the problem of multiple object tracking. Contemporary development in these fields brings them together, leading to the emergence of a unified model-based pattern recognition method. We formulate tracking as recognition of spatiotemporal patterns based on dynamic models, similar to image recognition based on geometric models.

Classical approaches to target tracking have been developed to track single targets and assume that an input to a tracking algorithm is comprised only of the target (position) data, whereas all extraneous or clutter data are perfectly rejected before tracking. Initial approaches to tracking were based on autoregressive models, also called Wiener filters in this application (Wiener, 1948) (fixed-coefficient filters, such as the once popular simple α–β tracker, are Wiener filters). A significant improvement in tracking came with the invention of Kalman filters that combined a complicated dynamic model of target motion with a probabilistic model of the observation process (Kalman, 1960). Classical tracking methods including Kalman filters considered data association as an afterthought.

In a Kalman filter formalism, a target track is characterized by a track model, by model parameters that are usually called state parameters, by model predictions of the expected values of data, and by covariances of the deviations between the data and predictions. We extend this to multiple targets. According to the notations of Section 1.2.5, we denote data vectors $\mathbf{x}_n$, their expected values or models $\mathbf{M}_k(\mathbf{S}_k)$ or simply $\mathbf{M}_k$, state or model parameters $\mathbf{S}_k$, and the hypothesis about the model and parameters H_k. In this section, a data vector corresponds to a single observation (not to a segment of data), and usually is comprised of coordinates. For example,

$$\mathbf{x}_n = (x_n, y_n) \tag{1.3-23}$$

where x_n and y_n are two angle coordinates or an angle and range coordinates. A data vector may also include other measurements, such as an amplitude of a signal, or several amplitudes for multiband sensors. An index n enumerates observations in multiple time frames, therefore there is a time parameter t_n associated with each observation n, which we will not usually show explicitly. The model includes the laws of motion, computing predicted data for every time point t_n; therefore, an index n is added to the model predictions, and models should be designed to predict expected values of the data for multiple points in time,

$$\mathbf{M}_{nk}(\mathbf{S}_k, t_n) = E\{\mathbf{x}_n | k, t_n\} \tag{1.3-24}$$

Consider a simple example of an unresolved target moving with a constant velocity $\mathbf{v}$. And let the data $\mathbf{x}_n$ be the Cartesian coordinates in some coordinate system. The state parameters completely characterizing the model in this case are

$$\mathbf{S}_k = (\mathbf{R}_k, \mathbf{V}_k) \tag{1.3-25}$$

where k is an index of a considered target and $\mathbf{R}_k$ is its initial position. The model is given by

$$\mathbf{M}_{nk} = \mathbf{R}_k + \mathbf{V}_k t_n \tag{1.3-26}$$

In this formulation, state parameters are constant whereas predicted data are propagated through time. Sometimes, it is convenient to propagate state parameters before computing observations for the following reasons. Equations of motion are often specified in a coordinate system different from the one in which the measurements are performed. Also, the number of measured coordinates can be less than have to be used in the equations of motion. For example, data from passive sensors such as a TV camera contain information only on angular positions of objects, whereas equations of motion might be given in 3-D Cartesian coordinates. Therefore, in addition to equations of motion, models often include coordinate transformations and can be nonlinear and significantly more complicated than (1.3-26). Also, equations of motion may include a rigid body rotation and other physical dynamic laws as appropriate.

Real-time implementation was a challenge for early tracking systems. Radar sensors provided very high rate data streams, whereas computers were not as powerful as those available today. Because of this, tracking algorithms had to be developed in a recursive form, that is, they utilized only the most recent data to update track parameters, while "forgetting" data as soon as possible. Kalman filters provide for this capability. However, this is achieved in Kalman filters at the expense of nonoptimal or simplified treatment of other aspects of the problem. There are many books devoted to Kalman filters. In this book we concentrate on development of more powerful methods, addressing the fundamental complexities of tracking that have not been addressed by Kalman filters. One of these complexities is the problem of association, which unifies tracking with pattern recognition.

1.3.5 Association Problem

The classical tracking methods described above have not considered the problem of multiple targets or clutter and noise rejection. When there is just a single target which is clearly visible in each data frame (or scan), for example, when a target signal is much stronger than all other signals, the detection step is trivial and is accomplished by thresholding the data. However, when multiple targets are present, or when the target signal is not much above noise or clutter, the problem of *association* of data and target tracks becomes the most complex aspect of tracking. Mathematically, the problem of target tracking in these complex cases resembles the problem of model-based vision. Detection and association in tracking correspond to the segmentation step in computer vision. Similarly, the need to associate multiple subsets of data with multiple possible tracks often leads to a combinatorial explosion of computational complexity.

Let us formulate the association problem within the model-based Bayesian framework. Then powerful probabilistic methods can be applied to the problem of data association. We will use the notations of Sections 1.2.3 and 1.2.5, with k standing for a particular track (moving object) and n enumerating measurements at several time points, t_n. Accordingly, the conditional pdf of the data $\mathbf{x}_n$ is given by

$$\text{pdf}\,(\mathbf{x}_n|H_k) = G\,(\mathbf{x}_n|\mathbf{M}_{nk}\,(\mathbf{S}_k, t_n)\,, \mathbf{C}_k) \tag{1.3-27}$$

Assuming a particular association between the data and models, Ξ, the conditional likelihood (conditioned on the association) is written similarly to (1.2-27),

$$L(\Xi) = \prod_k \prod_{n \in k} \text{pdf}\,(\mathbf{x}_n | H_k) \tag{1.3-28}$$

Here, $n \in k$ denotes pixels n associated with track k, and the likelihood is a product over all pixels, which is rearranged by tracks. The main difference in notations between (1.3-28) and (1.2-27) is that here each $\mathbf{x}_n$ is an individual data point (pixel) and not a segment (subset) of data points. Correspondingly here, the product over $n \in k$ corresponds to a segment (subset) of observations that is associated with the hypothesis H_k, and the product over k has the same meaning as in Eq. (1.2-27). Here, the joint pdf of all the pixels associated with the object k is taken as a product of individual pixels; again, as in Section 1.2, this does not require an assumption of their statistical independence: they can always be made statistically independent with the appropriate selection of models (see Note 8).

A solution to the association and track estimation problems is obtained by (1) maximizing the above expression over the set of track parameters, (2) repeating this step for all plausible associations between data and track hypotheses, and (3) selecting the association and track parameters corresponding to the maximum of the likelihood. This is similar to the model-based vision problem considered in Section 1.2.5, where the likelihood is maximized over parameters and segmentations. Similarly, a need to consider a combinatorially large number of associations leads to an explosion of computational complexity. Let us consider a concept used to reduce this complexity, which is popular in many association and tracking algorithms.

Association as Assignment. The Bayesian formulation for the association problem was developed by Sittler (1964); it became practically useful in the 1970s and 1980s after the widespread adoption of Kalman filtering techniques that provided for estimation of the covariance matrixes, $\mathbf{C}$. Historically, the problem of data association was considered after many years of successful development and practical implementation of recursive tracking algorithms. First formulations of the data association problem involved relatively few observations and tracks, for example, from crossing tracks, as illustrated in Fig. 1.3-2. In this problem, decisions have to be made as to which of the tracks has to be updated and which of the observations should be used. Within a recursive approach, this decision is made just once for each observation, using just the current frame (scan) of data. When tracks are well separated and clutter is not severe, association can be done by using track gates. A gate is a region around the predicted target position where observed data points are assigned to this target track. Gates are illustrated in Fig. 1.3-2; they can be specified by using track error covariances, for example, as $2-\sigma$ boundaries,

$$(\mathbf{x} - \mathbf{M}_k)^{\mathrm{T}}\ \mathbf{C}_k^{-1}\ (\mathbf{x} - \mathbf{M}_k) < 4 \tag{1.3-29}$$

In the presence of signals from noise, clutter, or closely spaced targets, the gates might overlap and data points may fall into more than one gate, so an approach is needed to make the best possible or most likely assignment. Using the likelihood expression (1.3-28), this problem can be formulated as the classical *assignment problem of linear programming*. The optimal assignment minimizes a total cost function, which is the sum of the individual assignment costs. Defining the individual assignment cost as the negative logarithm of pdf, (1.3-27), the total cost is given by the negative log-likelihood (1.3-28). Thus, the assignment problem is equivalent to maximizing the likelihood over various possible assignments. A

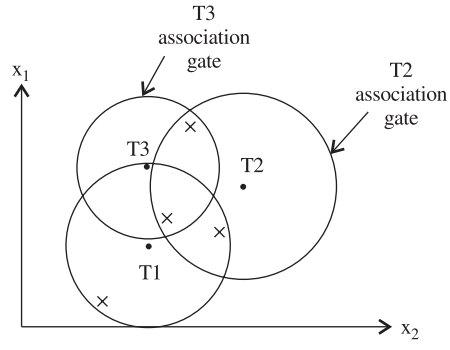


Figure 1.3-2 Examples of the data assignment problem; dots and circles show the predicted target positions and error bounds; crosses show measured data; error bounds are used as association gates.

typical formulation of the assignment problem involves first, computing a cost matrix, in our case $LP(n, k) = \ln \text{pdf}(\mathbf{x}_n | H_k)$, and then selecting a single entry from each column and each row so as to minimize the total cost, in our case the negative log-likelihood. An efficient algorithm for this problem was developed by Munkres (1957) and modified for tracking in Burgeois and Lassalle (1971). The computational complexity of this algorithm is on the order of n^2m, where $n = \min$ (number of observations, number of tracks), $m = \max$ (number of observations, number of tracks).

Assignment solves only one aspect of the problem; it assigns observations to tracks, *assuming that tracks have been accurately estimated*. Even the optimal solution of the assignment problem does not eliminate assignment errors, which, in turn, affect the accuracy of tracks. Errors in track estimation may lead to errors in gating and to more errors in data assignment, leading to lost tracks and poor performance. This problem has been partially addressed by considering misassociation as a source of errors that should be included into the Kalman filter estimation of the covariance matrices. When the number of possible misassociations is not very large, an assumption that tracks are known is appropriate, and the assignment technique can be used efficiently. However, in more complicated cases, when the number of misassociations and false alarms is large, initiating tracks becomes a most complex part of the problem. Tracking in imagery data, or estimating multiple concurrent dynamic processes in multidimensional data, cannot be solved by assignment algorithms, and concurrent association and track estimation are required.

Multiple Hypothesis Tracking. When target signals are of the same order of magnitude as clutter signals or when multiple targets are present, target detection cannot be performed on a single frame or scan. Multiple time measurements have to be utilized for target detection and association, which requires knowledge of tracks. Thus, the problems of detection, association, and track estimation have to be solved concurrently. The same is true for estimating multiple concurrent dynamic processes in multidimensional data and predicting such data. These problems resemble pattern recognition in space and time (or in multidimensional spaces that include time). Human and animal visual systems have separate subsystems that perform detection based on motion, that is, concurrent detection and tracking. Such a capability is referred to as track-before-detect or more accurately as track-while-detect. Solution of this problem requires associating multiple subsets of data with multiple possible track models, which often leads to a combinatorial explosion. Mathematical formulation and computational complexity of this general tracking problem resemble that of model-based vision. And similar mathematical concepts have been used

for attacking these problems. According to the concept of Multiple Hypothesis Tracking, the overall likelihood is maximized by a combinatorial search over a set of all possible track hypotheses. This concept is similar to the Multiple Hypothesis Testing approach discussed in Section 1.2.5 and we will use the same abbreviation, MHT.

In tracking applications, two approaches to forming hypotheses have been considered: first, hypotheses based on data partitions, and second, hypotheses based on partitioning the space of track model parameters. In the first approach, the data from multiple frames are partitioned in all plausible ways, similar to (1.2-26) (Fig. 1.3-3a). For each partition, the conditional likelihood given by (1.3-28) is maximized over parameters of all hypotheses, similar to (1.2-29). Due to factorization of the conditional likelihood (1.3-28) into the product of conditional track-likelihoods $[L(k, \Xi)]$, parameters of each track (conditional on segmentation Ξ) are estimated independently from other tracks,

$$\mathbf{S}_k(\Xi) = \arg\max L(k, \Xi); \quad L(k, \Xi) = \prod_{n \in k} \text{pdf}\,(\mathbf{x}_n | H_k) \tag{1.3-30}$$

The factorization greatly simplifies the parameter estimation problem, which now can be solved by using a Kalman filtering technique applied to a segmented data set $\{X_n, n \in k\}$ for estimating a single track k. After the best set of model parameters is obtained for every partition, the likelihood for every partition is computed and parameter values corresponding to the maximum likelihood partition are selected.

In the second approach, model parameter space is partitioned first (Fig. 1.3-3b). A set of hypothesis is formed, each comprised of a track model and specific values of the model parameters. Then the data are partitioned using an assignment algorithm, such as discussed in the previous section. For each hypothesis, track parameters are reestimated. At this point, the decisions are made as to which of the hypothesized tracks correspond to real targets and which do not. This can be done by applying likelihood ratio tests. To form a set of hypotheses based on partitioning the parameter space, it is necessary to decide on the coarseness of the hypotheses set in the space of parameter values. This is an important and nontrivial issue because too coarse a set of initial hypotheses may lead to lost tracks whereas too fine a set of initial hypotheses leads to a combinatorial explosion.

The two approaches to track hypotheses formation are inherently combinatorial, but their computational complexities are determined by different factors. The computational complexity, c, of the first approach, in which hypotheses are formed based on data partitions,

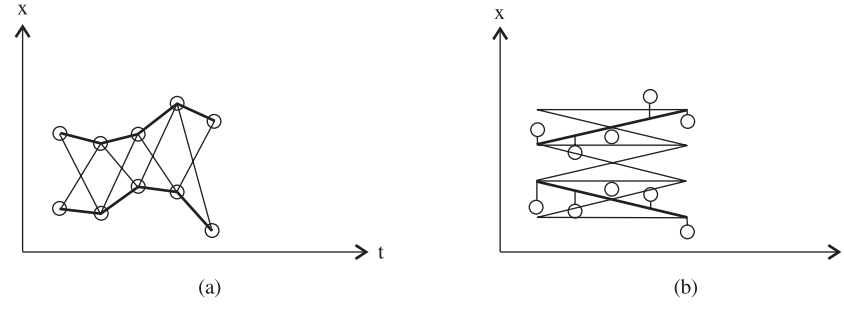


Figure 1.3-3 Association can be based on data partitioning **(a)** or parameter space partitioning models **(b)**. Data (○), association or track hypothesis (—), initiated tracks (▬).

is determined primarily by the number of data partitions. It grows combinatorially with the number of frames (time points) used in estimation, on the order of

$$c \sim C1^{*}k^{*}\mathrm{n}^{f} \tag{1.3-31}$$

where $C1$ is the number of operations required to maximize the conditional likelihood and estimate parameters for a single hypothesis, k is the number of tracks, n is the number of observations per frame (time point), and f is the number of frames.

Computational complexity of the second approach, in which hypotheses are formed by partitioning the parameter space, is determined primarily by the coarseness of model parameter specification. It grows combinatorially with the complexity of models, on the order of

$$c \sim C2^{*}k^{*}v^{p} \tag{1.3-32}$$

where $C2$ is the number of operations to compute data assignment, maximize the conditional likelihood, and estimate parameters for a single hypothesis, p is the number of parameters per model, and v is the number of values for each parameter, determined by the coarseness of the model specification. Factors $C1$ and $C2$ are approximately linear in the number of frames,

$$C1, C2 \sim f \tag{1.3-33}$$

The two tracking methods just discussed are modifications of the general MHT approach described in Section 1.2.5. The advantage of the MHT approach is that it is general and utilizes all information on multiple frames. This is especially important when target signals are weak relative to clutter signals so that targets cannot be reliably detected from a single frame, and multiple frames have to be utilized to determine the presence of a target. However, it is precisely under these conditions that MHT suffers from combinatorial explosion. When hypotheses are formed by data partition (the first approach), the combinatorial explosion is due to the large number of clutter returns on each scan and to the large number of frames that need to be used. When hypotheses are formed by partitioning the model parameter space (the second approach), the combinatorial explosion is due to the fine partition of the parameter space that is needed to achieve the required sensitivity. During the 1970s and 1980s several algorithms were developed based on this concept of maximizing the overall likelihood over the set of all possible tracks in multiple frames. These algorithms use different techniques to control the combinatorial explosion; still the combinatorial complexity is an essential property of the MHT concept related to considering combinations of many factors (data or parameters).

Probabilistic Data Association. A different tracking concept was proposed to overcome this essential combinatorial complexity of MHT algorithms. A concept of fuzzy association of tracks with multiple data points was developed by Bar-Shalom and co-workers (Bar-Shalom and Tse, 1975). For the case of a single target in clutter it was named Probabilistic Data Association (PDA), and for multiple targets in clutter, it was named Joint PDA (JPDA). In this technique, all measurements are probabilistically associated with each track using a posteriori Bayes probabilities, $P\,(H_k|X_n)$, (1.2-15). And the parameter update is computed by using a Kalman filter modified to account for this probabilistic association as

follows. In a standard application of the Kalman filtering technique, the parameter update is computed from the differences between the track measurement and its prediction (so-called residuals),

$$\mathbf{D}_{nk} = \mathbf{x}_n - \mathbf{M}_{nk}\,(\mathbf{S}_k, t_n) \tag{1.3-34}$$

In JPDA, the a posteriori probabilities are used as weights to compute a probability-weighted sum of residuals,

$$\langle \mathbf{D}_{nk} \rangle = \sum_n P\,(H_k|\mathbf{x}_n)\;\mathbf{D}_{nk} \tag{1.3-35}$$

And parameters of each track k are updated by using a Kalman filter with these weighted residuals. The JPDA concept relies on the existence of previously initiated tracks and performs parameter updates recursively, using only the data from the last scan. By using fuzzy data association, JPDA solved one part of the difficulty associated with the combinatorial explosion. However, JPDA relies on existent tracks. It is unsuitable for track initiation and cannot be used for target detection in heavy clutter nor for identification and prediction of temporal patterns in multidimensional signals. The problem of combinatorial explosion discussed in Section 1.2.5 for pattern recognition is equally relevant to tracking. This is not surprising when tracking is viewed as recognition of spatiotemporal patterns.

Conclusion. Target detection and track initiation in heavy clutter require concurrent processing of multiple frames (time points), while avoiding the combinatorial explosion. The same is true for identification and prediction of multidimensional dynamic patterns. The advantages of MHT and JPDA should be combined, while avoiding their drawbacks. Such a technique is considered in Chapter 7. Tracking multiple targets and identification and prediction of multidimensional temporal patterns are discussed there together with financial market prediction. We treat tracking multiple targets similarly to estimating concurrently operating market forces. Unknown tracks or market forces described by fuzzy adaptive models are similar to Aristotelian Forms, which in the process of real-time adaptation and learning become concepts of estimated tracks or market laws.

1.4 Preview: Intelligence, Internal Model, Symbol, Emotions, and Consciousness

This section previews the contents of the book from the vantage point of the first chapter. Chapter 2 continues reviewing mathematical concepts of intelligence. I attempted to overview most of fundamental mathematical concepts used for the modeling of mind. The domain is well beyond a single chapter. My attempt to overview this entire vast field is based on concentrating on the main *computational or mathematical* concepts. Because of space limitations, some of the mathematical concepts were left out of this review.[11] In the mathematical analysis, I had to leave out interesting variations and motivations, neural, psychological, cognitive, and many others. A mathematical analysis summarized in Chapter 2 shows that there are few basic "classical" computational concepts underlying most of algorithms of intelligence and neural networks. Each concept faces a combinatorial explosion of complexity, which became a nemesis of computational intelligence. Different

types of combinatorial complexity are related to the roles of a priori knowledge and adaptive learning. This analysis leads to close links between mathematical concepts of intellect and philosophical concepts of mind. Chapter 2 begins with the analysis of a contradiction between two concepts due to Aristotle: Aristotelian logic and Aristotelian theory of Form (theory of mind). And the ubiquitous problem of computational complexity is traced to this contradiction. Fuzzy logic is identified as a main ingredient needed to overcome combinatorial complexity.

Adaptive model-based fuzzy logic is discussed as a way to close the 2300-year gap between the logic and concepts of mind, to overcome mathematical difficulties, and to mend the schism between philosophy and mathematics. The stage is set for modeling field theory considered in Part II of the book. Then I review currently emerging computational concepts attempting to resolve the conundrum of combinatorial complexity: genetic algorithms, complex adaptive systems, mathematical semiotics, hierarchical organization, and neuronal fields. Relationships between these concepts and modeling field theory are discussed throughout the book.

Chapter 3 establishes detailed correspondence between the mathematical concepts of mind and philosophical concepts developed over the past 2300 years. Of course, this is a daring task, especially within the confines of a single chapter. I concentrate on those concepts that are most closely related to the current mathematical and scientific debates about the nature of mind. Although it is possible to argue that the entire human spiritual endeavor, including philosophy, gnosticism, alchemy, mysticism, and theology, is relevant to the scientific analysis of mind, much of this thought is still beyond mathematical analysis. Still, a surprisingly large area of thought from ancient Greek philosophers, to Gnostics, to Medieval philosophers and theologists, to Kant and post-Kantian development including psychological, cognitive, and mathematical debates of the nineteenth and twentieth centuries, evolves around a single issue, often called the "mind–body problem."

It might seem even more surprising that it is possible to trace continuous connections of concepts of mind in thinkers, separated by time, culture, and geography, along the lines of two main concepts. Materialism vs. idealism, realism vs. nominalism, immanence vs. transcendence, behaviorism vs. innateness, a priori knowledge vs. learning from experience, internal representations vs. situated behavior, parallel processing vs. sequential, neural vs. symbolic, and connectivism vs. logic are but a few names under which the proponents of the two main lines of thoughts draw the boundaries of their convictions. In the past, the debates might have led to spilled blood. In this century, it is usually research funding that is at stake. Throughout this book, I call these two main lines of thought *apriority* and *adaptivity* of mind. I have shown so far but a glimpse of the difficulty of unifying the two views.

Part II describes modeling field theory (MFT) and its engineering applications. MFT is a mathematical apparatus that combines apriority and adaptivity, and resolves the conundrum of combinatorial complexity by using model-based adaptive fuzzy logic. It consists of the three a priori ingredients: internal models that ascend to Plato and Aristotle, measures of similarity between internal models and input data, and the dynamic laws of adaptivity that maximize the similarity between the models and data, which comprises learning. Chapter 4 develops similarity measures and general laws of adaptive dynamics suitable for complicated internal models composed of multiple submodels. An architectural organization is discussed, including hierarchical and heterarchical types. Each submodel with its cooperative and competitive dynamics is identified as an intelligent agent within the overall system architecture.

Chapters 5 through 8 describe engineering applications and specific types of adaptive a priori internal models useful for these applications: classification and recognition; signal and image processing; spectrum estimation, including multidimensional time–space–frequency spectra; prediction and tracking, including complicated nonlinear relationships, multiple concurrent processes and targets, and noise and clutter; and sensor fusion, including data association and sensor management, which is related to the attention. These chapters include a number of examples of complex real-world problems, with the performance of the MFT significantly exceeding that of the prior state-of-the-art algorithms and neural networks.

Chapter 9 describes the fundamental mathematical limits on learning, depending on the amount of available data and contents of the a priori models. Part II concludes with Chapter 10, which continues the discussion of architectures of intelligent systems that we began in Section 1.1.4 with the discussion of the intelligent tracker. Chapter 10 establishes a close relationship between MFT and Kant's theory of mind, including emotions. The Kant–MFT theory of mind provides a foundation for the mathematics and physics of the concept of beauty. Then Chapter 10 discusses the relationships between MFT and complex adaptive systems, genetic algorithms, and semiotics, a science of signs and symbols. MFT agents are identified with the dynamic process of symbol formation, the process of semiosis combining internal representations, meaning, and behavior. Relationships among Kantian theory of mind, MFT, and semiotics are established.

Part III of the book discusses fun stuff: futuristic directions of research toward the physical theory of mind including consciousness, creativity, and free will. An important part of this investigation is delineation of what we can hope to understand from a rational scientific point of view, and what is currently beyond such hope. The line delineating boundaries of applicability of the scientific method is a moving one; still, it needs to be identified so that our scientific discussions could be properly focused. To this purpose, the first chapter of Part III (Chapter 11) starts with a discussion of the Gödel theorems concerning limitations of logic and their relevancy to the theory of mind. Recently, this topic provoked a heated debate involving physicists, mathematicians, and philosophers, with the final scores being far from settled. Gödel proved that formal systems related to Aristotelian logic or logic of predicates are fundamentally limited. Turing has reformulated this result for computational systems. There have been several attempts to use these results for proving the principled difference between the mind and machine. A most recent one is due to Penrose, who believes that the Gödel–Turing limitations have to be surpassed in order to model the mind. I briefly review the Gödel–Turing results, Penrose's arguments, and some counterarguments. I analyze the combinatorial nature of Aristotelian logic as revealed by the Gödel and Turing arguments, and compare it with the combinatorial explosion of complexity of intelligent algorithms and neural networks. My conclusion is that the Gödel–Turing results establish limitations to Aristotelian logic, but are not necessarily relevant to the theory of mind.

The last chapter, Chapter 12, is a prolegomenon to the future physical theory of consciousness. What is consciousness? Why is it needed in biological or artificial systems? Can it be understood as a physical phenomenon? Can it be described mathematically? I outline a future modeling field theory of consciousness. In this theory, consciousness is due to an internal model. The chapter overviews the phenomenology of consciousness and properties of the related internal model. It begins with popular conceptions and misconceptions and then continues the analysis, relating phenomenology to modeling field theory. A complex, differentiated nature of consciousness is discussed, and the phenomenology of

consciousness is described in its intimate connection to the rest of the psyche, including the unconscious and emotions. Hypotheses and historical evidence concerning origins and evolution of consciousness are summarized. Properties of consciousness are related to and explained within the modeling field theory. I overview neural structures involved in consciousness and emotions and identify candidate neural correlates for the modeling field theory modules and for the Kantian theory of mind.

The discussion continues toward more complicated aspects of consciousness, including the nature of creativity and free will. I analyze the differentiated nature of the process in which consciousness analyzes itself. This process is related to the nature of symbol in Jungian psychology and in modeling field theory. I identify an essential connection between creativity, consciousness, unconscious, and fuzziness, and attempt to delineate the boundaries of what is accessible today to the scientific method. Is it possible that mysteries of consciousness that are beyond rational understanding today are related to new physical phenomena, the discovery of which will resolve the mysteries of matter related to the yet unexplained nature of quantum measurement and quantum gravity?

NOTES

1. Marvin Minsky has contributed to a number of approaches to computational intelligence, including one of the very first neural networks (Minsky, 1954). He is most famous for his contribution to rule-based artificial intelligence (AI) or expert systems based on logical rules, which is referred to in our designation of the Plato–Minsky method. Rule-based AI is often called symbolic AI, because it operates with name variables rather than with numbers. I will usually avoid the symbolic AI designation, because it is a misnomer: name variables are signs, not symbols. Symbols are complex adaptive dynamic entities, of which name variables are only a small part. Mathematical description of symbols are considered in Chapter 10. See also Note 2.
2. In classical semiotics, words signs and symbols were not always used consistently. I designate a symbol as a dynamic process of concept formation and sign as a nonadaptive "mark." My designation is in line with analytic psychology and the general cultural usage of the word symbol as something having a profound effect on psyche.
3. These and other possible events are not alternatives, because they may or may not occur *in addition* to whichever would be the FRB decision. Our models for $\mathrm{pdf}(x|k)$, for $k = 1, \ldots, 4$, should account for these other possibilities to the extent that we can model them. Of course, our models are approximations; they could be inaccurate, and we should try to make them as accurate as feasible. But the set of alternatives always has to be complete, otherwise the theory is inconsistent. Sometimes it is difficult to come up with a complete set of alternatives. If this requirement is relaxed, one has to replace the theory of probability. For example, the Dempster–Shafer theory of evidence accumulation is a consistent theory dealing with random uncertainties without this requirement. In complicated cases, it might be faster to use the Dempster–Shafer theory than the theory of probability. But is it better in terms of performance? In my experience, when exact knowledge of all the circumstances is not known, still, using subjective sets of alternatives and adaptive models within the probability theory is preferable.
4. Calls and puts (also called options) are contracts traded at a stock exchange. A call gives its owner a right to buy a certain stock at a certain price (strike) before a certain day (expiration day). A put is a right to sell the stock at the strike price before the expiration day. Options are used both as highly risky investments and as a way of protecting assets against unexpected market moves.
5. The fundamental role of the Gaussian distribution in statistics is due to the fact that when uncertainty is caused by multiple random effects, the distribution most often is Gaussian. A theorem that proves this is called Central Limit Theorem (Cramer, 1946).

6. Estimation of the width of kernel functions (**C**) for the Parzen method is an unresolved problem for high dimensions: a narrow function requires too many training samples, whereas a wide function compromises sensitivity to interclass differences.
7. Definition of a sufficient statistics. Consider a pdf(data). Data, in general, is a set of many observations and/or dimensions, data $= (x_1, x_2, \ldots, x_N)$. For some functional shapes of the pdf, there is a single scalar combination of data, on which the pdf depends. If it exists, it is called a sufficient statistics. For example: $\text{pdf}(x_1, x_2, \ldots, x_N) = \sigma^{-N} \exp[-(x_1 + x_2 + \cdots + x_N)/\sigma]$. In this case, $f = (x_1 + x_2 + \cdots + x_N)$ is a sufficient statistics.
8. Note that the conditional pdf depend on class models and parameter values. This assumption is very different from the standard assumption of Gaussian densities. The models $\mathbf{M}_k$ that we consider can be of complex shapes in the data space (Chapter 7); therefore, the density (1.2-25) can model any deterministic variabilities in the data. In Chapter 5, we consider several types of non-Gaussian pdf including mixtures that can model arbitrary statistical variabilities.
9. To write the joint likelihood as a product, we do not need to assume independence of the data $\mathbf{x}_n$ and $\mathbf{x}_{n'}$ from different subsets ξ_n and $\xi_{n'}$. We treat as statistically independent the deviations from the data and models, $\mathbf{x}_n - \mathbf{M}_k$. These deviations can always be made independent with the appropriate definitions of the models. For example, let $\mathbf{x}_n$ and $\mathbf{x}_{n'}$ be statistically dependent, $\mathbf{x}_n = \mathbf{x}_{n'} + \varepsilon$, were ε is a random variable independent from $\mathbf{x}_n$ and $\mathbf{x}_{n'}$. Then, define the kth model for the pixel n as $\mathbf{M}_k = \mathbf{x}_{n'}$. This leads to $(\mathbf{x}_n - \mathbf{M}_k) = \varepsilon$ being statistically independent of $\mathbf{x}_{n'}$. This is further considered in Chapter 4, Problem 4.3-1.
10. The procedure used here of first deriving relationships that involve true theoretical values of parameters, and second, of substituting their estimated values is not grounded on a basic mathematical principle and thus is an ad hoc procedure. Sometimes it leads to an approximation of a more complicated, mathematically optimal expression. However, in the case of linear regression, we have shown that it is exactly *equivalent* to the optimal maximum likelihood estimation.
11. Among mathematical concepts that I would like to include in my review, given enough space, are chaotic attractors (Freeman, 1996), abductive logic (Kanal, in Bonnisone et al., 1991), inductive learning (Goldfarb, 1996), Hamiltonian dynamics of logic (Brockett, 1991), and my own work on quantum computation (Perlovsky, 1997c). I felt that some of these "left out" concepts are not crucially important to the main subject of this book, the physics of mind, others are related to the concepts already included in the review, and some were left out primarily because of space limitations. For example, a possibility of a chaotic dynamics of neural processes, which would be important for relating concepts of mind to the working of the brain, does not seem crucial to the theory of mind. At this point, a theory of mind is concerned with properties of the attractors, which are related to the modeling field theory developed in subsequent chapters. Still, given enough room I would have included these concepts. If a reader believes that there are other fundamental mathematical concepts relevant to the theory of mind that I left out, please let me know.

BIBLIOGRAPHICAL NOTES

This section contains references that were not directly given in the text, as well as some additional ones. The order here is approximately historical.

Antisthenes, a founder of the Cynic school of philosophy, lived in Athens in fifth century BC. His ideas were preserved in the works of other Greek philosophers (further see Windelband, 1893).

Plato's theory of ideas was developed throughout his life; it's essential features are discussed in Parmenides (IV BC); and Timaeus (IV BC).

Aristotelian theory of Forms (theory of mind) is discussed in Aristotle's Metaphysics (VI BC) and On Psyche (VI BC). Aristotle developed a theory of logic in several works, including Prior Analytics, Posteriori Analytics, Topics (see Aristotle, VI BC, *The Complete Works of Aristotle*).

Probability theory, Bayes theory, and estimation theory: for further reading see, e.g., Cramer (1946) and Anderson (1984). An alternative to probability estimation is called Statistical Learning Theory (Vapnik, 1995). I borrowed two concepts from Vapnik: referring to the probability theory as solving the forward problem and to statistics as solving the inverse problem.

Fuzzy logic (Zadeh, 1962, 1965, 1997).

McCulloch: modeling of the neural structures of the brain (McCulloch and Pitts, 1943) and the influence of Occam vs. the realistic logic (McCulloch, 1961, 1965).

Early artificial neural network: see Minsky and Papert (1988) and further references therein; Rosenblatt (1958; 1962).

For the foundations of expert or rule systems: see Minsky (1968a. 1975).

Rule systems and learning, in general (Minsky, 1975; Winston, 1984), in pattern recognition (Winston, 1984; Bonnisone et al., 1991; Keshavan et al., 1993), and in linguistics (Koster and May, 1981; Botha, 1991).

Chomsky's linguistics (Chomsky, 1972, 1981; Botha, 1991). The last book contains an excellent and lively review of the field up to 1991.

Combinatorial complexity (Bellman, 1961; Winston, 1984; Segre, 1992; Perlovsky, 1994a).

Pattern recognition (Nilsson, 1965; Fukunaga, 1972; Duda and Hart, 1973; Watanabe, 1985). Classical techniques of statistical pattern recognition are summarized (Fukunaga, 1991, 2nd edition). Feature construction by approximating sufficient statistics (Perlovsky et al., 1992a).

Model-based approaches to machine vision (Nevatia and Binford, 1977; Brooks, 1983; Winston, 1984; Grimson and Lozano-Perez, 1984; Chen and Dyer, 1986; Lamdan and Wolfson, 1988; Negahdaripour and Jain, 1991; Bonnisone et al., 1991; Segre, 1992; Keshavan et al., 1993; Califano and Mohan, 1994). National Science Foundation symposium on machine learning (Negahdaripour and Jain, 1991).

Tracking algorithms: classical "Hungarian" assignment algorithms (Munkres, 1957), modified for tracking (Burgeois and Lassalle, 1971; Marty, 1976); association errors included into the Kalman filter estimation of the covariance matrices (Nahi, 1969; Jaffer and Bar-Shalom, 1972); MHT tracking algorithms (Singer et al., 1974; Reid, 1979; Blackman, 1986); Probabilistic Data Association tracking algorithms, PDA and JPDA (Bar-Shalom and Tse, 1975; Fortmann et al, 1980).

Mathematical analysis of basic computational concepts of intelligence (Perlovsky, 1994a; 1998a; also see Winston, 1984; Simpson, 1990; Girosi et al., 1995).

PROBLEMS

1.1–1 Establish a correspondence between the Aristotelian concept of Form, semiotical concepts of sign, designatum, interpretant, and the intelligent tracker operation. *Hint:* identify material entities in the world, a priori contents of mind, and results of mind's adaptation to the world. Note that the Aristotelian and semiotical concepts describe various aspects of the intelligent tracker, which only partially overlap. The Aristotelian description concentrates on the role of the a priori and on the process of adaptation. The semiotical description emphasizes the structural elements, but ignores the mechanism of the adaptation.

1.2–1 Prove that $G(\mathbf{x}) = \text{constant}$ is the equation of an ellipsoid in D-dimensional $\mathbf{x}$-space. *Hint:* remember, that an ellipsoid is given by a second-order equation in terms of $x_1, \ldots, x_D$. *Comment:* a second-order equation could also correspond to a parabolic or hyperbolic curve; in case of Gaussian density, this curve is always an ellipse due to the positive definiteness of a covariance matrix. Further explanations could be found in Searle (1982).

1.2–2 Prove that for a Gaussian $\text{pdf}(\mathbf{x}) = G(\mathbf{x}|\mathbf{M}, \mathbf{C})$,

$$\mathbf{M} = E\{\mathbf{x}\}, \text{ and } \mathbf{C} = E\{(\mathbf{x} - \mathbf{M})(\mathbf{x} - \mathbf{M})^{\mathrm{T}}\}$$

Note: for the beginner, it is sufficient to understand that the above expressions are the basis for the identification of the Gaussian pdf parameters with the corresponding expected values. An

advanced student should go through the following, as it teaches important tricks for deriving more complicated theoretical relationships throughout the book and in mathematical statistics and linear algebra, in general. (1) Prove the first of the above equations by observing that $\int(\mathbf{x}-\mathbf{M})G(\mathbf{x}|\mathbf{M}, \mathbf{C})\,d\mathbf{x} = 0$, because $(\mathbf{x}-\mathbf{M})$ is an asymmetrical function about $(\mathbf{x}-\mathbf{M}) = 0$ and $G(\mathbf{x}|\mathbf{M}, \mathbf{C})$ is a symmetrical one. (2) Prove the second of the above equations by (2.1) using the equality $\int G(\mathbf{x}|\mathbf{M}, \mathbf{C})\,d\mathbf{x} = 1$ and (2.2) taking the derivative, $\partial/\partial(\mathbf{C}^{-1})$; in the one-dimensional case, evaluate this derivative straightforwardly; in the multidimensional case, use the following identities that are true for any matrix $\mathbf{C}$, $(\det \mathbf{C})^{-1} = \det(\mathbf{C}^{-1})$, $\partial/\partial\mathbf{C}(\det \mathbf{C}) = \det \mathbf{C}/\mathbf{C}^{-1}$, $\partial/\partial(\mathbf{C}^{-1})(\mathbf{D}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{D}) = \mathbf{D}\mathbf{D}^{\mathrm{T}}$.

1.2–3 Formulate Eq. (1.2-12) in terms of probabilities, rather than pdfs, so that it will match Eq. (1.2-2) exactly. *Hint:* multiply each side of this equation by $d\mathbf{x}$.

1.2–4 Formulate Eq. (1.2-13) in terms of probabilities, rather than pdfs. *Hint:* multiply the numerator and denominator on the right-hand side of this equation by $d\mathbf{x}$.

Comment. Is $P(H_k|\mathbf{x})$ a probability or a pdf? In other words, when the accuracy of our knowledge of $\mathbf{x}$ is very high, so that $d\mathbf{x} \to 0$, shouldn't $P(H_k|\mathbf{x}) \to 0$? The arguments on this issue can be presented as follows. If $\mathbf{x}$ is known within a certain finite-size region (over which the a priori pdf vary substantially), the size and shape of the region have to affect the a posteriori probabilities of various hypotheses $P(H_k|\mathbf{x})$. But when the size of this region is so small that the pdf do not vary over this region, the exact accuracy of the measurement of $\mathbf{x}$ should not affect our decisions concerning probabilities of various hypotheses.

1.2–5 Verify Eq. (1.2-21). *Hints:* $-0.3q-0.3(1-p-q) = -0.3(1-p)$; $(1.2p-0.3) > (1.2q-0.3)$ is equivalent to $p > q$, and $(1.2p - 0.3) > 0$ is equivalent to $p > 0.23$.

1.3–1 Solve the Regression Equations. Obtain (1.3-11 and 1.3-12) from (1.3-8). *Hint:*

1. Multiply the second of Eqs. (1.3-8) by $\bar{x}$, and subtract it from the first of Eqs. (1.3-8). Obtain

$$\sum_n (y_n - ax_n - b)\ (x_n - \bar{x}) = 0$$

2. Using the first of Eqs. (1.3-9) prove that

$$\sum_n (\bar{y} - a\bar{x} - b)\ (x_n - \bar{x}) = 0$$

3. Subtract this equation from the first of Eqs. (1.3-8) and obtain Eqs. (1.3-11).
4. Obtain Eqs. (1.3-12) straightforwardly from (1.3-9).

1.3–2 Normalized Variables. Instead of the data set $\{(x_n, y_n)\}$, consider a data set $\{(x'_n, y'_n)\}$, with $x'_n = (x_n - \bar{x})$, $y'_n = (y_n - \bar{y})$. Following the procedure in Section 1.3.1, rederive a linear regression equation $y'_n = ax'_n$ and prove that it is equivalent to Eq. (1.3-1) with coefficients given by (1.3-13). *Lesson:* it is more convenient to use normalized zero-mean variables (x', y') than the original variables (x, y).

Repeat the above with $x'_n = (x_n - \bar{x})/\sigma_x$, $y'_n = (y_n - \bar{y})/\sigma_y$. *Lesson:* it is even more convenient to use normalized zero-mean unit-standard-deviation variables (x', y') than the original variables (x, y).

1.3–3 Write a computer code to solve the linear regression problem, Eqs. (1.3-13). Select any values for a and b, any small (positive and negative) values for ε_n , and fill the bottom row of the

following table. Run your computer code, estimate a and b, and compare the estimated and true values ($y_n = ax_n + b$). Predict the y value for $x = 11$ and compare it with the true value. Plot the data, and the estimated and true regression lines.

$x_n =$	1	2	3	4	5	6	7	8	9	10
$y_n = ax_n + b + \varepsilon_n$										

1.3–4 Exercise in Manipulating Gaussian Densities. For the Gaussian densities given by (1.2-4) and (1.2-9), verify the general relationship $\text{pdf}(x) = \int \text{pdf}(x, y)\, dy$. *Solution:*
We use the following notations: σ_x and σ_y are standard deviations of x and y and r is the correlation coefficient between x and y; covariance matrix, $\mathbf{C} = \begin{Bmatrix} C_{xx} & C_{xy} \\ C_{xy} & C_{yy} \end{Bmatrix}$; elements of this matrix are $C_{xx} = \sigma_x^2$, $C_{yy} = \sigma_y^2$, $C_{xy} = r\,\sigma_x\,\sigma_y$. The determinant and inverse of matrix $\mathbf{C}$ are given by

$$\det \mathbf{C} = C_{xx}\, C_{yy} - C_{xy}^2 = \sigma_x^2\, \sigma_y^2 \left(1 - r^2\right) \tag{P1.3-4a}$$

$$\mathbf{C}^{-1} = \begin{Bmatrix} C_{yy} & -C_{xy} \\ -C_{xy} & C_{xx} \end{Bmatrix} \Big/ \det \mathbf{C} \tag{P1.3-4b}$$

For shortness, we will use normalized variables but will denote them with the same letters x and y:

$$x \text{ and } y \text{ instead of } (x - \bar{x}) \text{ and } (y - \bar{y}) \tag{P1.3-4c}$$

With these notations, Gaussian densities for x and (x, y) are written as

$$\text{pdf}(x) = \left(2\pi\sigma_x^2\right)^{-1/2} \exp\left(-0.5x^2/\sigma_x^2\right) \tag{P1.3-4d}$$

$$\text{pdf}(x, y) = (2\pi)^{-1} (\det \mathbf{C})^{-1/2} \exp\left(-0.5\, \mathbf{D}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{D}\right) \tag{P1.3-4e}$$

$$-0.5\, \mathbf{D}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{D} = -0.5 \left[(x/\sigma_x)^2 + \left(y/\sigma_y\right)^2 - 2r\, xy/\left(\sigma_x \sigma_y\right)\right] \Big/ \left(1 - r^2\right) \tag{P1.3-4f}$$

Let us rearrange the exponent in Eq. (P1.3-4f) as

$$\begin{aligned} &-0.5 \left[(x/\sigma_x)^2 + \left(y/\sigma_y\right)^2 - 2r\, xy/\left(\sigma_x \sigma_y\right)\right] \Big/ \left(1 - r^2\right) \\ &= -0.5\, (x/\sigma_x)^2 - 0.5\, \left(y - r\, x\, \sigma_y/\sigma_x\right)^2 \Big/ \left[\sigma_y^2\, \left(1 - r^2\right)\right] \end{aligned} \tag{P1.3-4g}$$

Now, let us compute

$$\begin{aligned} \int \text{pdf}(x, y)\, dy &= \int (2\pi)^{-1} (\det \mathbf{C})^{-1/2} \exp\left(-0.5\, \mathbf{D}^{\mathrm{T}} \mathbf{C}^{-1} \mathbf{D}\right)\, dy \\ &= \left(2\pi\sigma_x^2\right)^{-1/2} \exp\left(-0.5\, x^2/\sigma_x^2\right) \\ &\cdot \int \left[2\pi\sigma_y^2\, \left(1 - r^2\right)\right]^{-1/2} \\ &\exp\left[-0.5\, \left(y - r\, x\, \sigma_y/\sigma_x\right)^2 \Big/ \left(\sigma_y^2\, \left(1 - r^2\right)\right)\right] dy \end{aligned} \tag{P1.3-4h}$$

Comparing the last line here to Eq. (P1.3-4d), we observe that the integrand is the Gaussian density for variable y, with the

$$\text{mean} = \left(r\, x\, \sigma_y/\sigma_x\right) \text{ and standard deviation } = \left[\sigma_y^2\ \left(1 - r^2\right)\right] \qquad \text{(P1.3-4i)}$$

The Gaussian density with any mean and standard deviation is normalized so that its integral equals 1. The first line here equals (P1.3-4d), which completes the proof.

1.3–5 Regression as Expectation. For the Gaussian densities given by (1.2-4) and (1.2-9), derive (1.3-18). Outline of the solution: using (1.3-17) and (P1.3-4d) and (P1.3-4h), show that the last line in (P1.3-4h) is $\text{pdf}(y|x)$. According to (1.3-16), the expected value of y given x is the mean value of $\text{pdf}(y|x)$, and is given by (P1.3-4i). Remember that (P1.3-4i) is written for the normalized variables, (P1.3-4c); substituting the original variables, obtain (1.3-18).

1.3–6 Stock Market Data. Go on the Internet and find stock market data (e.g., America-On-Line has free data for several major indexes, such as Dow Jones, SP-500, etc.). Select any 12-month interval. Evaluate performance of the most simple strategy "buy and hold": buy on day 1, hold for 12 months, sell after 12 months.

1.3–7 Stock Market Simplest Prediction. Write two computer codes (1) for prediction and (2) for trading evaluation based on predictions. Start with simple codes:

1. Simple prediction code: tomorrow's change equals today's change; that is, if on day n the market is up, that is, $\text{DJ}_n > \text{DJ}_{n-1}$, the prediction is $+1$; if $\text{DJ}_n < \text{DJ}_{n-1}$, the prediction is -1.
2. Simple trading evaluation code: start with \$1 cash. If the prediction is > 0, buy using all your cash (if any, otherwise do nothing). If the prediction is < 0, sell all your holdings (if any, otherwise do nothing). At the end of the 12-month interval, compute your gain (or loss). Compare it to "buy and hold" strategy in Problem 1.3-6. Hopefully, the prediction-based strategy is not worse than "buy and hold." (In a bear market, even the simple prediction strategy would be better; in a bull market, it is very difficult to predict better than the market does on average.)

1.3–8 Autoregressive Prediction. Select 1 month of data prior to the 12 months considered above and estimate the autoregressive coefficient Eqs. (1.2-11, 1.3-9, 1.3-10, 1.3-15, 1.3-22) from this month data (training data). Replace your prediction code in Problem 1.3-7 with an autoregression prediction, Eq. (1.3-22). Evaluate performance over 12 months and compare with the Problem 1.3-7 result. Continue with the next two problems.

1.3–9 The Stationarity Assumption. The stationarity assumption can be verified by comparing statistics over various time intervals. Select a 12-month interval of your data. Compute statistics, $\bar{x}$, r, and σ, 12 times separately over each of the 12 months.

1.3–10 Autoregressive Prediction using Differences. Repeat Problem 1.3-8 using $x'_t = x_t - x_{t-1}$. Compare results to Problem 1.3-7. If there is an appreciable difference (beyond computer accuracy and round-offs), then examine your autoregressive coefficient, r: it should be a small positive number. If $r > /0$, then (1) is it ≤ 0 within your computer roundoff errors?; (2) did you select a training interval over which $r \leq 0$? (see Problem 1.3-9); (3) look for bugs in the code.

1.3–11 Adaptive Autoregressive Prediction. Repeat Problem 1.3-10 using $x'_t = x_t - x_{t-1}$. Recompute the autoregression coefficient r every day using the past month of data. Compare results to Problem 1.3-7. If there is an appreciable difference (beyond computer accuracy and round-offs), then examine your autoregressive coefficient, r. If $r > /0$, then (1) is it ≤ 0 within your computer roundoff errors?; (2) does r really change sign? (see Problem 1.3-9); if so, you should do better here than in Problem 1.3-7; (3) look for bugs in the code.

MATHEMATICAL CONCEPTS OF MIND

> Aristotle, I heard you are writing books now. Are you going to make our secret knowledge public?
>
> —FROM A LETTER BY ALEXANDER
>
> Alexander, do not worry: nobody will understand.
>
> —FROM A REPLY LETTER BY ARISTOTLE

This chapter overviews computational concepts of intellect. A ubiquitous problem facing intelligent algorithms and neural networks is their exploding computational complexity and training requirements. Although Aristotelian logic is still the basis for most of our algorithms, an argument is formulated that a solution to the conundrum of combinatorial complexity will have to utilize fuzzy logic. This argument is traced through the rest of the chapter.

We begin with a mathematical analysis of the four basic computational concepts used in traditional approaches: the nearest neighbor, gradient learning, rule-based, and parametric model-based concepts. Each concept faces a combinatorial explosion of computational complexity, which became the nemesis of computational intelligence. Different types of complexity are related to the roles of a priori knowledge and adaptive learning. This establishes a connection between mathematical concepts of intellect and concepts of mind discussed by philosophers since Plato. Difficulty of combining apriority and adaptivity was first analyzed and philosophically resolved by Aristotle in his theory of Form. We trace contemporary mathematical difficulties involving complexity to the contradiction between Aristotelian theory of Form and Aristotelian logic. Adaptive model-based fuzzy logic is discussed as a way to close the 2300-year gap between logic and concepts of mind, to overcome mathematical difficulties, and to mend the schism between philosophy and mathematics. After reviewing classical mathematical concepts and their deficiencies, we turn to currently emerging mathematical concepts proposed to overcome the combinatorial computational complexity: genetic algorithms and evolutionary computation, mathematical semiotics, hierarchical organization, and neural fields. The stage is set for modeling field theory considered in the second part of the book.

Many excellent books describe in detail popular neural networks and production systems. This chapter, instead, analyzes basic computational concepts underlying specific rule-based and neural paradigms, so that one can foresee their general capabilities and limitations.

Section 2.1 provides a conceptual preview for the chapter. Sections 2.2 through 2.5 describe classical computational concepts of intelligence. Sections 2.6 through 2.9 address

"transitional" techniques that were designed in search for resolving the conundrum of combinatorial complexity. In particular, Section 2.8 discusses major debate issues and controversies, including the mathematical nature of thought, understanding, and emotions. And, finally, Sections 2.10 through 2.14 describe newly emerging concepts having a potential to overcome the conundrum of combinatorial complexity.

2.1 Complexity, Aristotle, and Fuzzy Logic

2.1.1 Conundrum of Combinatorial Complexity

Intelligent computational methods and applications in diverse areas have faced difficulties that are of related origin. We have already touched on this subject in Chapter 1. In areas of adaptive control, pattern recognition, artificial intelligence, machine vision, image understanding, information and sensor fusion, solution of complex problems have met with difficulties that are expressed in terms of the complexity of the solution process. Various computational paradigms have their own sets of difficulties, but it seems that there always is a step in the solution process that is exponentially or combinatorially complex.[1] A well-known term used in this regard is "the curse of dimensionality" (Bellman, 1961), which refers to a combinatorial increase in the required number of training samples with the increase of the dimensionality of a control or a pattern recognition problem. A similar problem is encountered in function approximation, and other areas (Girosi et al., 1995). The curse of dimensionality is characteristical of adaptive algorithms and neural networks that learn from data.

Another set of difficulties is encountered by computational paradigms that utilize systems of a priori logical rules—rule-based artificial intelligence or rule-based AI. In the case of rule systems, the difficulty is in the fast (combinatorial) growth of the number of rules with the complexity of the problem (Winston, 1984). Model-based approaches that utilize process or object models in the control or recognition process encounter difficulties manifested as combinatorial complexity of required computations (Nevatia and Binford, 1977; Brooks, 1983; Grimson and Lozano-Perez, 1984). The difficulties of various paradigms have been summarized in recent reviews as follows. "Much of our current models and methodologies do not seem to scale out of limited 'toy' domains" (Negahdaripour and Jain, 1991). "The key issues [are] . . . the inherent uncertainty of data measurements" and "combinatorial explosion inherent in the problem" (Grimson and Huttenlocher, 1991).

The seemingly inexorable combinatorial explosion that reincarnates in every intelligent computational paradigm is related in this chapter to the fundamental issue of the roles of a priori knowledge vs. adaptive learning. This relationship was discussed in Perlovsky (1994a; 1998a) for geometric patterns and in Girosi et al. (1995) for function approximation. The issue of the roles of a priori knowledge vs. adaptive learning has been of an overriding concern in the research of mathematics of intelligence since its inception. In the introduction to Chapter 1, we briefly discussed the controversy about a priori knowledge and learning tracing it throughout the entire history of the concepts of mind through the Middle Ages to Aristotle and Plato and established links between philosophy and mathematics. The

philosophical thoughts of the past turned out to be directly relevant to the development of mathematical concepts of intellect today. The next section discusses mathematical difficulties of relying either on adaptive learning or on a priori knowledge (which is discussed in more details in later sections). The discussion then turns to difficulties of combining apriority with adaptivity in the presence of uncertainty. Arguments are presented for Aristotelian logic being culpable for the combinatorial complexity of combining adaptivity and apriority.[2] We discuss adaptive fuzzy logic as an approach to reducing algorithmic complexities. This section (1) previews classical computational concepts of intelligence and summarizes more detailed discussions of Sections 2.2 through 2.5 and (2) provides a basis for discussing newly emerging concepts in Sections 2.9 through 2.15.

2.1.2 Adaptivity, Apriority, and Complexity

McCulloch, staying at the cradle of mathematical modeling of neural networks, believed that the material basis of the mind is in complicated neural structures of a priori origin. Specialized, genetically inherited a priori structures have to provide for specific types of learning and adaptation abilities. An example investigated by McCulloch was a group-averaging structure providing for scale-independent recognition of objects, which McCulloch believed serves as a material basis for *concepts* or *ideas* of objects independent of their apparent size (Pitts and McCulloch, 1947). This direction of inquiry, however, was not continued during the early neural network research in the 1950s and 1960s. Most notable neural networks developed at that time, Perceptron and Adaline, utilized simple structures. Their design concept was general learning from data without a priori knowledge. These neural networks were the first examples of learning machines inspiring a generation of researchers; still their capabilities were fundamentally limited.

In the 1960s and 1970s, pattern recognition algorithms were developed utilizing the concept of classification or decision space (Nilsson, 1965; Fukunaga, 1972; Duda and Hart, 1973; Watanabe, 1985). Their concept also was general learning from data without a priori knowledge. Pattern recognition algorithms often faced difficulties related to combinatorial training requirements. Exorbitant training requirements of statistical pattern recognition and control algorithms can be understood due to the geometry of high-dimensional classification spaces. There are three basic approaches used to partition a classification space into class or decision regions: model-based parametric classifiers, nearest neighbors, and gradient learning using superpositions of planar surfaces (Perlovsky, 1994a). Simple decision regions could be defined by using parametric models of their boundaries, such as derived from Gaussian distributions. Like early neural networks, they are fundamentally limited to simple shapes (such as the quadratic classifier). Nonparametric paradigms (gradient learning and nearest neighbors) have been used to surpass limitations of simple parametric methods. However, due to the fact that the volume of a classification space grows exponentially with the dimensionality (number of features), training requirements for nonparametric paradigms are often exponential in terms of the problem complexity (Perlovsky, 1994a). This is essentially the same problem that was encountered earlier in the field of adaptive control and was named "the curse of dimensionality" (Bellman, 1961). The father of cybernetics, Wiener, also acknowledged this problem, he emphasized that using higher order predictive models, or combining many simple models, is inadequate for the description of complex nonstationary systems, because of insufficient data for learning (Wiener, 1948). Mathematical

analysis of nonparametric paradigms and their learning difficulties is presented in Sections 2.2 and 2.3. The algorithms designed for learning from data are shown to come up against the fundamental limitation of combinatorial growth of learning requirements.

To avoid learning difficulties, Minsky suggested a different concept of artificial intelligence (AI) based on the principle of apriority. He argued that intelligence could be understood only on the basis of extensive systems of a priori rules (Minsky, 1968a). This was the next attempt (after McCulloch) to understand the intellect from the principle of apriority. The main advantage of this method is that it requires no training, because it explicitly incorporates detailed, high-level a priori knowledge into the decision-making process. The a priori knowledge is represented in a set of logical rules similar to the high-level cognitive concepts utilized by a human in conscious decision-making processes. However, systems of logical rules turned out to be brittle in changing environments that did not exactly fit into the design categories of a particular system. More and more rules should be added to account for variabilities, leading to a combinatorial explosion of the complexity of rule systems. A natural approach to overcoming this difficulty seemed to be to add adaptivity to the rule systems. However, combining adaptive learning with a priori knowledge proved difficult: variabilities and uncertainties in data required more and more detailed rules leading to combinatorial complexity of logical inference (Winston, 1984). Section 2.4 is devoted to analyzing this paradigm.

Model-based approaches to machine vision have been used to extend the rule-based concept to 2-D and 3-D sensory data. Physically based models enable utilization of detailed a priori information on objects' properties and shape in algorithms of image recognition and understanding (Nevatia and Binford, 1977; Brooks, 1983; Winston, 1984; Grimson and Lozano-Perez, 1984; Chen and Dyer, 1986; Michalski et al., 1986; Lamdan and Wolfson, 1988; Negahdaripour and Jain, 1991; Bonnisone et al., 1991; Segre, 1992; Keshavan et al., 1993; Califano and Mohan, 1994). Models used in machine vision typically are complicated geometric 3-D models that require no adaptation. These models are useful in applications in which variabilities are limited and types of objects and other parameters of the recognition problem are constrained. When unforeseen variabilities are a constant factor in the recognition problem, utilization of such models involves difficulties that are common to rule systems. More and more detailed models are required, potentially leading to a combinatorial explosion.

Parametric model-based algorithms have been proposed to overcome the difficulties of previously used methods by combining the adaptivity of parameters with the apriority of models. In these approaches, adaptive parameters are used to adapt models to the variabilities and uncertainties in data. Parametric adaptive methods date back to Widrow's Adaline and linear classifiers. These early parametric methods can be efficiently trained using few samples; however, they are limited to simple control laws and simple decision regions and are not suitable for complex problems. Complicated problems, such as image recognition, require utilization of multiple flexible models. In the process of recognition, an algorithm has to decide which subset of data corresponds to which model. The available data set of pixels or samples has to be subdivided into subsets corresponding to classes or models. This step is called segmentation, or association, and it requires a consideration of *multiple combinations* of subsets of the data. Because of this, complicated adaptive models often lead to combinatorial explosion of the complexity of the solution. Section 2.5 is devoted to analyzing a model-based paradigm.

The preceding discussion can be summarized as follows. A mathematical analysis of classical approaches to the design of systems and algorithms of mathematical intelligence leads to a conclusion that *computational* concepts of most of today's neural networks and fuzzy systems originate in classical control and pattern recognition algorithms, and that there are four basic concepts forming the foundation for all the multiplicity of classical algorithms and neural networks (Perlovsky, 1994a; 1998a). These are (1) the concept of nearest neighbors and (2) the concept of gradient learning, both defined by the factor of adaptivity (Nilsson, 1965; Duda and Hart, 1973); (3) the concept of rule systems (Minsky, 1968), defined by the factor of apriority; and (4) the concept of parametric models (Winston, 1984; Segre, 1992), which attempts to combine apriority and adaptivity. Whereas methods based on adaptivity face combinatorial explosion of *the training process*, those based on apriority face combinatorial explosion of *the complexity of rule systems,* and attempts to combine the two face combinatorial explosion of *computational complexity.* Factors of apriority and adaptivity ought to be combined by physically acceptable concepts of the intellect. Therefore, approaches to combining both factors are of paramount interest. However, existing computational paradigms have not resolved the problem of combinatorial complexity. To repeat again, "Much of our current models and methodologies do not seem to scale out of limited 'toy' domains" (Negahdaripour and Jain, 1991); "The key issues [are] . . . the inherent uncertainty of data measurements" and "combinatorial explosion inherent in the problem" (Grimson and Huttenlocher, 1991).

2.1.3 Fuzzy Logic and Complexity

Fuzzy logic can play a crucial role in reducing computational complexity of model-based approaches to combining adaptivity and apriority, because it eliminates the continuum that causes combinatorial complexity. The precision of Aristotelian logic allows for the continuum of concepts: another concept can always be "inserted between" however similar Aristotelian concepts, while still preserving its own unique identity. Thus, Aristotelian concepts that form only countable sets cannot exhaust "all possible Aristotelian concepts," leading to logical contradictions such as the Russell paradox and Gödel theorem of incompleteness. This line of reasoning is further pursued in Chapter 11; here we illustrate how fuzzy concepts eliminate combinatorial complexity of Aristotelian concepts using a simple example.

Consider a problem of segmentation or association, which leads to combinatorial explosion of image recognition or nonlinear prediction. The segmentation or association problem is illustrated in Fig. 2.1-1 using crisp membership functions for a one-dimensional problem. Selecting the best association between the data $\{X\}$ and concepts-models (Fig. 2.1-1) during the process of adaptation requires evaluation of combinatorially many subdivisions.[3] The combinatorial complexity is illustrated for more complicated two-dimensional cases in Fig. 2.1-2. The combinatorial complexity of Aristotelian logic is related to the high (absolute) precision of every concept, which may not be warranted by an approximate nature of data or knowledge.

The combinatorial complexity can be avoided by using fuzzy associations as illustrated in Fig. 2.1-3a. Here, every data point is associated with many concept-models and combinatorial explosion is avoided. In addition, the fuzzy association in Fig. 2.1-3a accounts for possible uncertainties in the data, which may not warrant the crisp memberships of Figs. 2.1-1 and 2.1-2. The price paid, however, is a degree of fuzziness leading to overlap

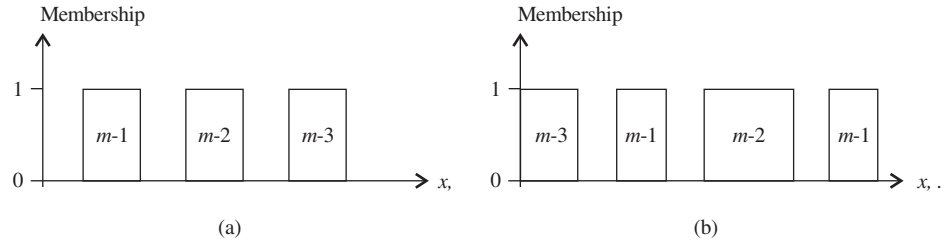


Figure 2.1-1 Combinatorial complexity of a segmentation or association problem is illustrated using crisp membership functions for a one-dimensional case (X is a feature or a measurement). Two associations are shown out of all possible combinations between models ($m-1, m-2, m-3, \ldots$) and subsets of X.

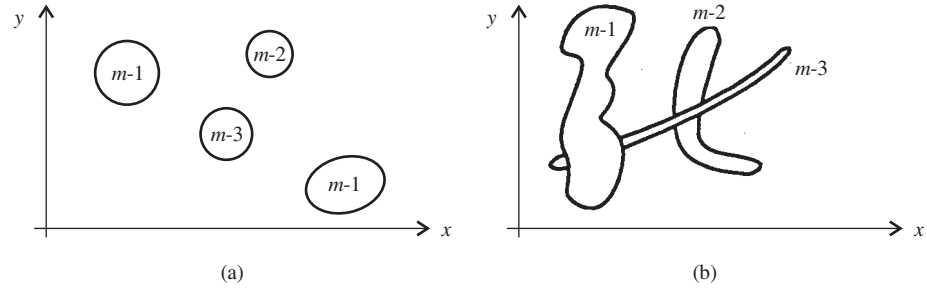


Figure 2.1-2 Association problem is illustrated for two-dimensional cases; (**a**) a simple case, (**b**) a more complicated case. Class boundaries correspond to the boundaries of the crisp membership functions.

between classes (models), which could be too large, unwarranted, and unacceptable. The founder of fuzzy logic, Zadeh, emphasized that a fixed resolution of fuzzy concepts imposes a limitation on usefulness of fuzzy logic for intelligent systems. An improvement can be achieved by using adaptive membership functions, so that in the process of adaptation, the initial association in Fig. 2.1-3a evolves into Fig. 2.1-3b and c. It is further illustrated for a two-dimensional case in Fig. 2.1-4. This procedure, although contradicting the Aristotelian logic, closely follows the Aristotelian description of the learning process, in which Form-as-potentiality becomes a concept. Such a procedure can be effected by combining existing fuzzy logic and adaptive clustering techniques. What needs to be developed is a combination of adaptive and fuzzy mathematics with the concept of apriority, utilizing a priori adaptive models as illustrated in Fig. 2.1-5. Here, an initial highly fuzzy configuration (a) evolves into a final low-fuzzy configuration (b), which is determined by both the data and the adaptive parametric models of these shapes. A combination of flexible model-based mathematics with adaptive fuzzy mathematics has not before been available. Development of such adaptive fuzzy model-based mathematics that I call modeling field theory or model-based neural network is the subject of this book and is pursued in detail in Chapters 4 through 7. The rest of this chapter is devoted to analysis of classical and emerging computational concepts of modeling intelligence.

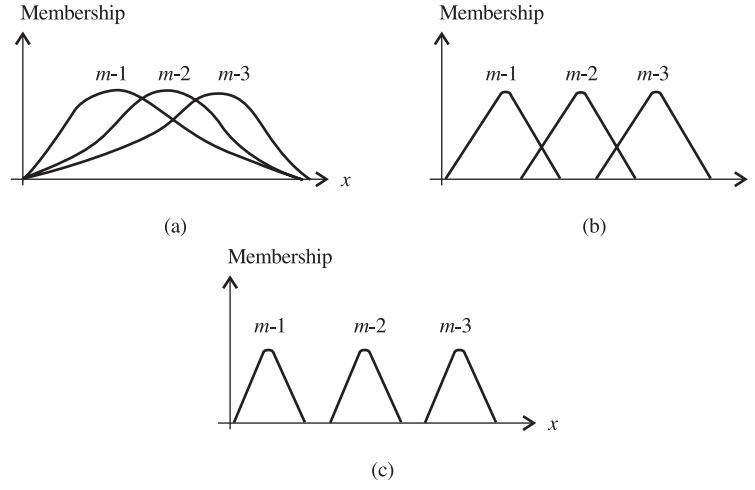


Figure 2.1-3 Adaptive fuzzy association avoids combinatorial explosion.

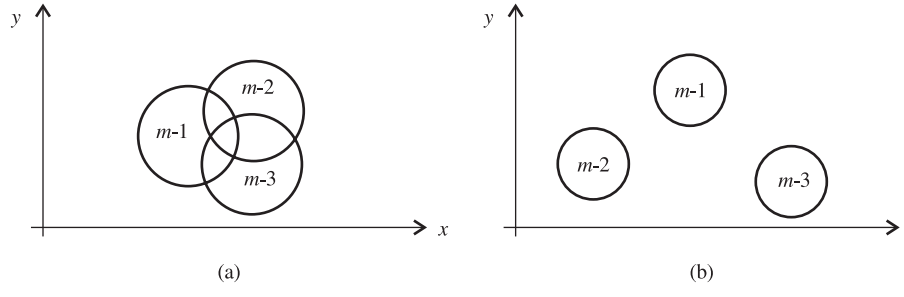


Figure 2.1-4 Adaptive fuzzy association in a two-dimensional case; (**a**) initial, (**b**) after adaptation. Class boundaries are indicated by 0.5-membership contours of fuzzy membership functions.

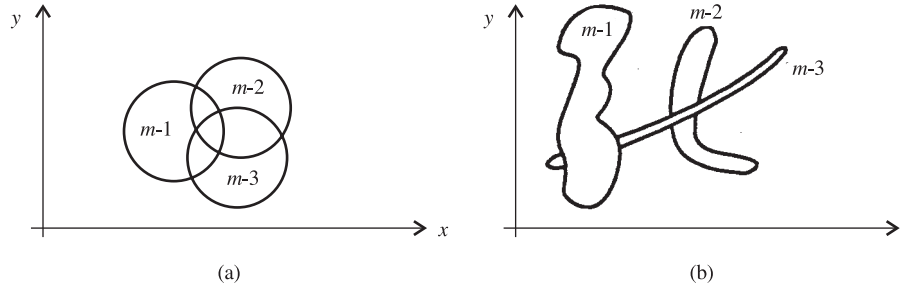


Figure 2.1-5 Adaptive model-based fuzzy association; (**a**) initial, (**b**) after adaptation.

2.2 Nearest Neighbors and Degenerate Geometries

Classification and recognition of objects are among central problems of computational intelligence. In this section we begin the analysis of the intelligent algorithms with the nearest neighbor concept of recognition.

2.2.1 The Nearest Neighbor Concept

An obvious and straightforward classification concept is to classify each sample to the same class or concept as the nearest (the most alike) sample from past experience (training sample). The nearest neighbor concept (NNC) is the simplest mathematical realization of the nominalistic concept of intellect, according to which *ideas* and *concepts* emerge in the process of learning from experience as names of classes of similar objects (and not from a priori knowledge). It is a highly intuitive concept and it serves as a basis for a large number of algorithms and neural networks. Samples are considered as vectors in a metric *classification space*, their coordinates are defined by measured signals or calculated feature values, and a Euclidean metric is most often used to calculate distances between new samples, their neighbors, and classes or concepts. The algorithms differ in selecting initial concepts and in deciding which training samples to retain in the memory as representatives of classes.

The name *nearest neighbor* is used for algorithms that retain in their memory all or many of the exemplars defining each class and arrive at a classification of a newcoming sample by calculating its distance from each of the exemplars in the memory. For high-dimensional objects, such as visual images or speech signals, the memory requirement can grow quickly; it is necessary, therefore, to limit the number of stored samples. Mathematical algorithms of many neural paradigms and their relationships to the NNC have been analyzed by Simpson (1990). For example, the Additive neural network often called the Hopfield net (McCulloch and Pitts, 1943; Grossberg, 1968; Hopfield, 1982) averages samples from each class during training to estimate and store means for each class. During classification, each sample is classified to a class according to the nearest mean; therefore the name of these algorithms is *the nearest mean* algorithms.

The NNC can be used for both supervised and unsupervised learning. For the latter case, various approaches are utilized to define new classes, to merge old classes, and to update means of the classes, as in the Competitive learning networks (Grossberg, 1970; Kohonen, 1984). Other neural networks based on the nearest neighbor concept include Reduced Coulomb Energy (RCE) (Reilly et al., 1987), Probabilistic Neural Network (PNN) (Specht, 1990), Neocognitron (Fukushima and Miyake, 1982), and Regularization neural networks (Poggio, 1988).

There is a fundamental limitation to the NNC in high dimensional classification spaces, leading to exorbitant training requirements. The NNC requires a dense sampling of the decision space with training examples, in turn, requiring a combinatorially large number of training examples, NT, as a function of dimensionality of the decision space D, $\mathrm{NT} \sim 10^D$ (Perlovsky, 1994a). Even medium size problems of visual imagery or speech recognition require considerations of hundreds of dimensions, $D > 100$, and the corresponding training requirements for the NNC exceed the number of objects in the universe.[4] This physically unsatisfiable requirement points to the principal inadequacy of the nominalistic concept of mind—learning from examples without using a priori knowledge is physically impossible.

2.2.2 Mathematical Formulation

A mathematical formulation of the nearest neighbor recognition concept is as follows. Figure 2.2-1 illustrates classification of each sample to the same class as the nearest training sample. Each sample $\mathbf{x}$ is considered as a vector in a metric classification space $\mathbf{x} = (x_1, \ldots, x_D)$, its coordinates x_i could be sensor measurements or feature values calculated from measurements. Figure 2.2-1 shows two of N coordinates. A Euclidean metric for calculation of distance $d(\mathbf{x}, \mathbf{y})$ between two samples $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$d^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{D} (x_i - y_i)^2 \tag{2.2-1}$$

A simplest nearest neighbor recognition algorithm can be specified as follows: during training, store all training samples $\mathbf{x}_n, n = 1, \ldots, N$, together with their class labels, $k(n)$. When a new sample $\mathbf{y}$ is coming, find the distances between $\mathbf{y}$ and all $\mathbf{x}_n, d(\mathbf{x}_n, \mathbf{y})$. Then, find the minimal distance $d(\mathbf{x}_{n-\min}, \mathbf{y})$, and classify y to class $k(n - \min)$.

2.2.3 What Constitutes Simple and Complex Classification Problems?

Before turning to the fundamental limitation of the nearest neighbor concept in high dimensional classification spaces, let us consider two types of classification problems illustrated in Fig. 2.2-2.

In a simple case (a), the two classes are relatively far from each other, and any classification approach easily separates these two classes. A nearest neighbor algorithm would need only one example from each class to learn perfect classification without errors. And there will be no difficulty with a similar case in a high-dimensional space. The complex case in (b) is different: if one tries to extend this case to high dimensions, one will find that the nearest neighbor concept is completely useless. In Section 2.2.2, we will explain why

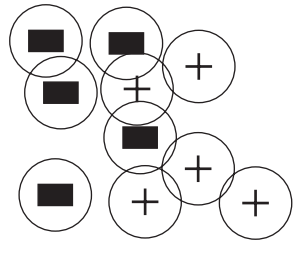
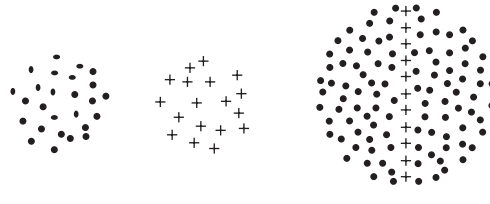


Figure 2.2-1 The nearest neighbor classification concept.



(a) (b)

Figure 2.2-2 Two examples of a two-class problem: (**a**) simple geometry: far separated classes, (**b**) complicated 'gridlock' geometry: two classes are well separated, but not "far" from each other.

this example is important. We will demonstrate that this type of classification geometry, resembling a gridlock in high dimensions, is common for image and signal classification and for many other complex recognition problems. Although in this case the two classes are well separated, the nearest neighbor approach will have to learn enough training samples to provide for a dense sampling of the classification space. We will show now that for high-dimensional spaces this requirement is prohibitive.

Consider the one-dimensional uniform distribution in Fig. 2.2-3a. If n samples are drawn from this distribution randomly, an average distance between neighboring samples will be $d = n^{-1}$. A similar result for two dimensions (b) will be $d \sim n^{-1/2}$, and for D dimensions $d \sim n^{-1/D}$. We see that in high dimensions the average distance between neighboring samples remains almost unchanged as the number of samples increases. To make distance between neighboring samples significantly smaller than the size of a class as a whole, we will need on the order of $n \sim 10^D$ samples. In the example of Fig. 2.2-3, to make samples from each class closer to each other than to the samples from the other class, more and more samples need be retained in the memory for higher dimensions, say, for $D \sim 100$ dimensions, it will be necessary to retain on the order of $n \sim 10^{100}$ samples. An order of magnitude of this estimate is independent of the shape or uniformity of the distribution: learning by the nearest neighbor rule requires a dense sampling of the decision space and this, in turn, requires an exponentially large number of samples.

The inadequacy of the nearest neighbor concept in high-dimensional classification samples has been observed by many experts in the field of pattern recognition. However, in the absence of the formal proof given above, and also, because the nearest neighbor concept is efficient for simple problems and in low-dimensional spaces, algorithms and neural networks based on the nearest neighbor concept remain popular. They can be used for simple problems, but one should not be surprised when a solution does not scale up for more complicated problems.

2.2.4 Degenerate Geometry of Classification Spaces

In this section we will demonstrate that a gridlock-type geometry of classification space illustrated in Fig. 2.2-2b is not an exception, but is commonplace for complicated problems. Recognition of objects, for example of image patterns, is based on the interrelationship

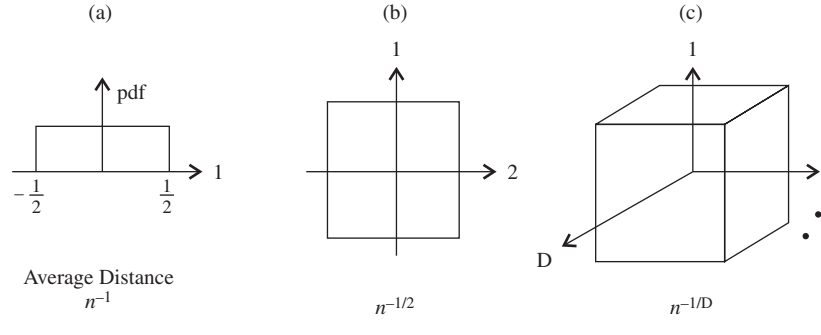


Figure 2.2-3 Relationship between dimensionality and near neighborliness.

between the object's parts: for a solid object in three-dimensional space, given the positions of any three points, all the other points of the object are fixed. This trivial observation leads to a nontrivial conclusion: there are high correlations between dimensionalities in pixel space; if one would like to train an algorithm or a neural network to classify objects in pixel space (without sophisticated procedures for feature extraction), one will often face degenerate geometries of the type shown in Fig. 2.2-2b.

This is illustrated in Figs. 2.2-4 and 2.2-5. Figure 2.2-4 shows two simple geometric patterns, a vertical bar and a horizontal bar, in a 3×3 pixel image. For translation invariant recognition, there are only three different patterns of each bar in a binary representation. For classification in pixel space each pattern is coded as a nine-dimensional vector of zeros and ones. Thus, classification space is nine-dimensional, with each axis being an intensity of a corresponding pixel.

The geometry of classification space is shown in Fig. 2.2-5 for the two-dimensional subspace of the first two components (the first two pixels in the upper row of a 3×3 image). The vertical-bar class occupies three points in this subspace, (0,0), (0,1), and (1,0). The horizontal-bar class occupies only two points, (0,0) and (1,1). If gray-scale invariance

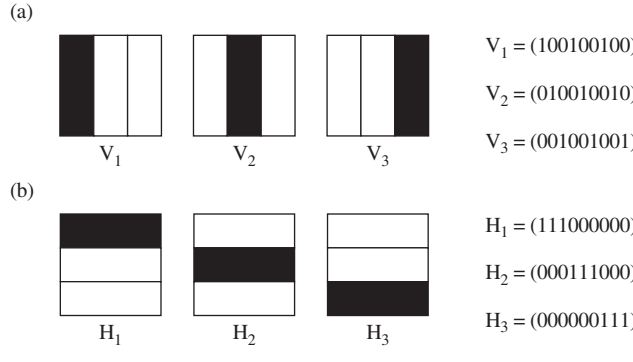


Figure 2.2-4 Three images. (**a**) There are three images corresponding to a vertical bar. (**b**) There are three images corresponding to a horizontal bar.

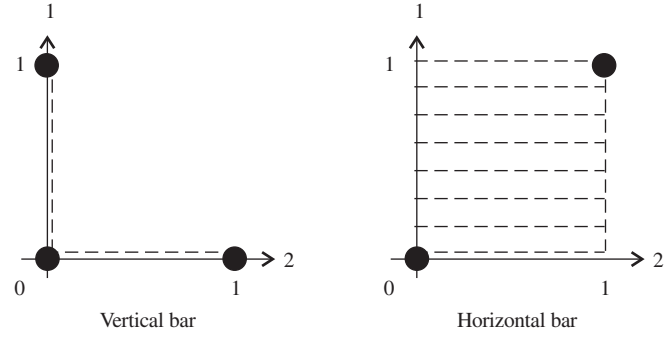


Figure 2.2-5 The classification space of the two classes of Fig. 2.2-4 (only the first two dimensions—pixels are shown). These classes, like a multidimensional gridlock, occupy close regions in classification space and cannot be effectively separated by a nearest neighbor classifier, yet they are completely separable.

is desirable (illuminance insensitivity), so that the gray scale can vary between 0 and 1 for each bar-pixel independently, the vertical-bar class occupies both axes between 0 and 1 and the horizontal-bar class occupies the whole dashed square. Thus, in this subspace the vertical-bar class occupies only a zero volume; in other subspaces a horizontal-bar class occupies only a zero volume.

These two classes occupy the same general area of the classification space, still they can be separated without error due to their degeneracy: the overlap between the two classes is 0% of the total volume of each class. To separate these two classes without an error using the nearest neighbor classifier, it is necessary to retain in the memory an infinite number of samples from each class. If a small error is permitted, a finite (large) number of samples will suffice, providing for a dense coverage of the classification space. As discussed in Section 2.2.2, a dense coverage in high dimensions requires a combinatorially large number of samples rendering the nearest neighbor concept useless.

Does the human visual system rely on the nearest neighbor learning of patterns in the world? The answer is no. We know very well that our visual system (as well as vision of other animals) relies on a priori models: edge recognition, moving dot detection, and other simple-pattern recognition are based on inborn models. These models are encoded in the structure of ganglion cells, which determine the a priori properties of the receptive fields in the retina. For simple patterns, these models are hardwired so that almost no adaptation is possible or needed. More complex functions of the visual system (e.g., stereoscopic vision) are genetically prewired in such a way that some adaptation is possible (and has to occur for successful functioning). As far as we know, the nearest neighbor recognition is not employed by the visual system.

Let us summarize. The nearest neighbor concept is intuitive and practically useful for simple problems. However, complex problems with degenerate geometry in high-dimensional classification spaces cause difficulties for the nearest neighbor approaches. A combinatorially large number of training samples is needed to learn what is similar and what is dissimilar. The nearest neighbor concept is the simplest straightforward mathematical implementation of the nominalistic conception of mind. According to the nominalistic philosophy, all the knowledge is acquired from experience, no a priori ideas are needed, and ideas and concepts are conventional names of classes of similar objects. Difficulties of the nearest neighbor concept are related to the general deficiency of the nominalistic concept: there are no measures of similarity in the world, they cannot be "just learned from experience," and they have to be inborn, a priori. The philosophical concept of nominalism cannot explain working of the mind.

2.3 Gradient Learning, Back Propagation, and Feedforward Neural Networks

2.3.1 Concept of Discriminating Surfaces and Gradient Learning

Whereas in the nearest neighbor concept an algorithm remembers *volume* of the multidimensional classification space in the neighborhood of training samples, another mathematical realization of the nominalistic concept of intellect consists in remembering a *boundary*

separating classes or concepts in the classification space. Discriminating surfaces is a concept according to which learning is a search for a collection of planar surfaces making up a classification boundary. This concept was originally explored in pattern recognition research in the 1960s (i.e., Duda and Fossum, 1966; Ho and Agrawala, 1968; Specht, 1967; Nilsson, 1965), and today this concept is revived in multilayer feedforward neural networks or multilayer perceptrons and in several other similar architectures (e.g., recurrent networks). The backpropagation mechanism most often used with these neural networks is based on the gradient descent for learning parameters of these surfaces; therefore we will call this concept "gradient learning of discrimination surfaces." Backpropagation was designed for supervised training and is not capable of self-learning within the internal neural dynamics. However, the analysis below concentrates on more fundamental limitations of this type of network, limitations related to the nominalistic concept in general.

Feedforward neural networks have been analyzed in detail by many authors (Moore and Poggio, 1988; Carpenter, 1989). We summarize results of these analyses in terms of classification space geometry and then analyze the mathematical reasons for the slow learning of this type of neural network. Longstaff and Cross (1987) related a number of layers in a feedforward neural network to its capability of representing a classifier of a complicated shape and explained the role of nonlinearity in multilayer networks. A standard neuron is shown to partition the classification space with a hyperplane into two halves and, therefore, to be capable of representing a linear classifier. The multineuron, single-layer network is capable of representing multiple linear classifiers, and a nonlinear transformation of the neuronal output does not modify the nature of this neural network, remaining a collection of linear classifiers.

A two-layer network is capable of representing a nonlinear classifier boundary, while a nonlinear transformation is essential—otherwise the second layer of neurons would still be computing linear combinations of the original input. With the nonlinear transformation, it is easy to verify that each second-layer neuron is capable of representing an inside region of any convex shape delineated by hyperplanes computed at the first layer. The exact shape of the nonlinear transformation is not essential for the representational purpose: a smooth function results in smooth boundaries. However, the implementation of a backpropagation learning mechanism, which uses gradients, requires a smooth sigmoidal function. The third layer of neurons is capable of combining regions computed at the second layer, similar to the second layer combining hyperplanes computed at the first layer. By combining a number of convex regions, a three-layer feedforward network is capable of representing a classifier boundary of arbitrary nonlinear shape.

Experience with feedforward neural networks has shown that they are widely applicable to relatively simple problems; however, they are poorly scaled up to large, real-world problems. Even though there are examples of constructing complicated, structured networks based on a feedforward architecture (Lang et al., 1990), they are based on extensive experience and require a significant developmental effort for each concrete problem. In essence, a priori knowledge of each problem has to be built into the feedforward neural architecture—but feedforward networks have not been designed for and do not facilitate such a procedure. Learning, thus, has to occur *in spite of* the architectural concept of these neural networks designed for the purpose of general learning requiring no a priori knowledge, based on the nominalistic concept of learning from examples.

The main impediment to application of these neural networks is their slow learning: a

huge number of training examples is required even when a special architecture is designed for a concrete problem. Mathematical analysis of this learning disability shows that in a general case of a D-dimensional classification space, a feedforward neural network requires at least an order of $20 \cdot D^2$ training examples for each deterministically differentiable subclass—a convex region in a classification space that does not overlap with other classes (Perlovsky, 1994a). For complicated but practically important cases, it is necessary to consider spaces of hundreds of dimensions and the number of subclasses (while smaller than the required number of training samples in the nearest neighbor concept) may increase combinatorially, thus rendering feedforward neural networks unsuitable. These neural networks cannot fulfill an expectation of the adaptive model of the intellect, self-learning from examples, without utilizing a priori knowledge—one more mathematical example that nominalism (learning from experience without a priori knowledge) is not a physically acceptable concept of mind.

2.3.2 Mathematical Formulation

The concept of gradient learning is illustrated in Fig. 2.3-1: learning is a search for a collection of planar surfaces making up a classification boundary.

The beginning of our analysis follows Longstaff and Cross (1987) explaining a relationship between the number of layers, a neural network's capability to represent a classifier of a complicated shape, and explaining the role of nonlinearity in multilayer networks. Consider a typical neuronal operation,

$$y = (\mathbf{wx}) + w_0 = \sum_{i=0}^{D} w_i x_i \tag{2.3-1}$$

where $\mathbf{w} = (w_1, \ldots, w_D)$ are neuronal weights, $\mathbf{x} = (x_1, \ldots, x_D)$ are input (for notation convenience define $x_0 = 1$), w_0 is a threshold ($x_0 = 1$ is defined for notation convenience), and y is an output (before a nonlinear transformation). Equation (2.3-1) defines a hyperplane in $\mathbf{x}$-space, such that $y > 0$ on one side of the hyperplane and $y < 0$ on the other side. It follows that a neuron calculating the sign of the expression (2.3-1) is partitioning the classification space with a hyperplane into two halves, and is therefore capable of

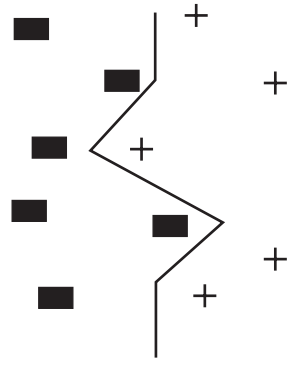


Figure 2.3-1 Discriminating surface concept.

representing a linear classifier (Fig. 2.3-2a). The multineuron single-layer network of Fig. 2.3-2b is capable of representing multiple linear classifiers. Note that if a nonlinear transformation is applied to the neuronal output y, $O = s(y)$, were s is a sign or a sigmoidal function, this does not modify the nature of the neural network: it is still a collection of linear classifiers.

A two-layer network, as represented in Fig. 2.3-3a, is capable of representing a nonlinear classifier boundary. At this step, it is necessary to introduce a nonlinear transformation, $o = s(y)$, otherwise the second layer of neurons would still be computing linear combinations of the original input $\mathbf{x}$. With the nonlinear transformation, it is easy to verify that the second-layer neurons are capable of representing the inside region of any convex shape delineated by hyperplanes computed at the first layer. The exact shape of the nonlinear transformation is not essential for the representation purpose: a smooth function results in smooth boundaries. However, for the implementation of a backpropagation learning mechanism (as well as any gradient method), a smooth sigmoidal function is important.

The third layer of neurons, Fig. 2.3-3b, is capable of combining regions computed at the second layer, similar to the second layer combining hyperplane-defined regions computed at the first layer. The three-layer feedforward network is therefore capable of representing a classifier boundary of arbitrary shape.

Learning consists in defining parameters of Eq. (2.3-1), $w_0, \ldots, w_D$, for each neuron. These parameters are defined so that the error between the desired output and actual output of the network is minimized. To define the error function and derive the learning equations, let us first introduce appropriate notations for multilayer feedforward networks. Consider a network with two layers of neurons (Fig. 2.3-3a). It has input signal nodes, middle (or hidden) layer nodes, and output nodes; three indexes i, j, k are used for each layer of nodes. The lower, input signal nodes contain the components of the data pattern (or vector) x_i. The weighted sums of these input data at the middle (or hidden) layer are denoted y_j, etc., as defined in Table 2.3-1.

The error function, on presentation of the input pattern $X_n = (x_{n1}, \ldots, x_{nD})$ is defined as

$$E_n = 0.5 \sum_k (T_{nk} - O_{nk})^2 \tag{2.3-2}$$

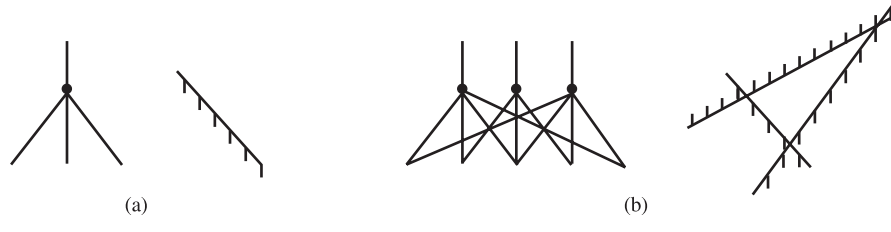


Figure 2.3-2 A single neuron (**a**) is capable of representing a linear classifier. A single-layer network (**b**) is capable of representing a collection of linear classifiers.

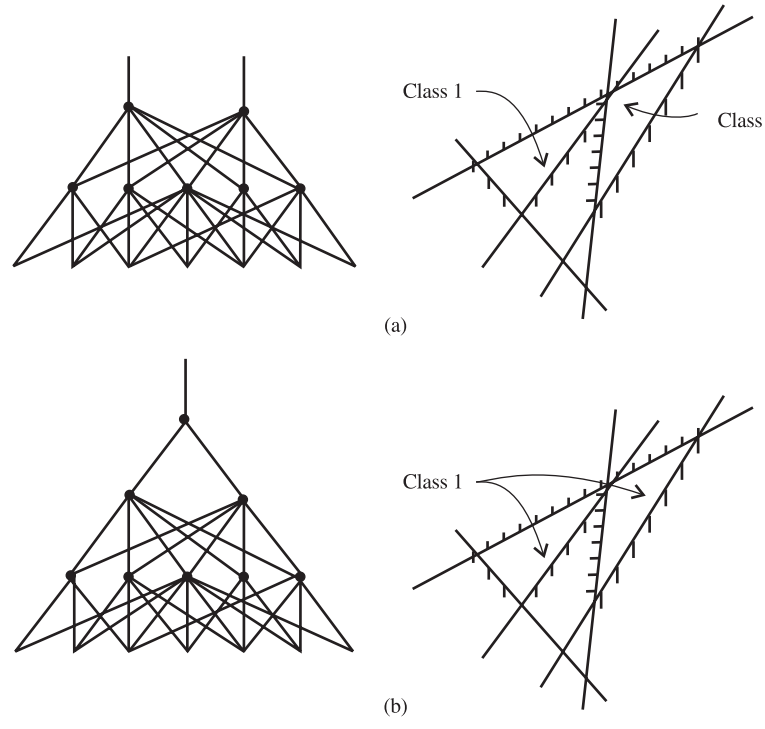


Figure 2.3-3 A two-layer network (**a**) is capable of representing convex-shaped classes. A three-layer network (**b**) is capable of representing classifier boundaries of any shape.

Using the definitions in Table 2.3-1, the explicit expression for the network output O_{nk} is

$$O_{\mathrm{nk}} = s\left(y_{nk}\right) = s\left(\sum_j w_{kj} O_{nj}\right) = s\left[\sum_j w_{kj} s\left(y_{nj}\right)\right]$$
$$= s\left[\sum_j w_{kj} s\left(\sum_i w_{ji} x_{ni}\right)\right] \tag{2.3-3}$$

The learning equations, specifying the weight update after presentation of each pattern, are determined by the gradient descent, $\Delta w = -\eta \partial E_n / \partial w$, and, by combining Eqs. (2.3-2) and (2.3-3), we obtain the explicit equations called error backpropagation (Werbos, 1974; Parker, 1985; Rumelhart et al., 1986):

$$\Delta_n w_{kj} = -\eta \partial E_n / \partial w_{kj} = \eta \left(T_{nk} - O_{nk}\right) s'\left(y_{nk}\right) O_{nj}$$
$$\Delta_n w_{ji} = -\eta \partial E_n / \partial w_{ji} = \eta \sum_k \left(T_{nk} - O_{nk}\right)\ s'\left(y_{nk}\right) s'\left(y_{nj}\right) w_{kj} O_{ni} \tag{2.3-4}$$

Here, s' denotes the derivative of the nonlinear transform s, and η is a constant determining the convergence rate.

TABLE 2.3-1
Definition of a Feedforward Neural Network with Two Layers of Neurons

Layer	Index	Input	Weights	Weighted Sum	Output	Target Output
Output	k	$O_{nj} = s(y_{nj})$	w_{kj}	$y_{nk} = \sum_{i=0}^{D} w_{kj} O_{nj}$	$O_{nk} = s(y_{nk})$	T_{nk}
Middle (hidden)	j	x_{ni}	w_{ji}	$y_{nj} = \sum_{i=0}^{D} w_{ji} x_{ni}$	$O_{nj} = s(y_{nj})$	
Input	i	x_{ni}				

2.3.3 Learning Disability

Experience with feedforward neural networks has shown that they are general and applicable to a variety of problems. However, they are poorly scaled up to large, real-world problems; in particular, they are notoriously slow learners. Three types of learning problems have been reported in the literature: first, the large number of iterations needed for convergence, given the training data set; second, convergence to local minima of the error function; third, the large number of training samples required for learning. The first two types of problems can be mitigated to some extent by modifying the learning algorithm. The third one is due to the fundamental limitation of the nominalistic concept of learning from examples without a priori knowledge.

Here we analyze only the third problem. The analysis of degenerate geometries of classification spaces in Section 2.2.3 showed that a large number of hyperplanes might be needed to separate classes in complicated cases. Examine Fig. 2.2-5: to separate the two classes shown there, a single hyperplane is needed for each of the two shown dimensions, so this problem is easy to solve with feedforward neural networks. However, consider a different problem: shift-invariant recognition of classes defined by checkerboard-type patterns with boxes of varying size. These types of classes occupy corners of a D-dimensional cube in the pixel-classification space. Even a few classes of patterns can be defined so that their invariant recognition requires checking a majority of cube corners. In terms of the feedforward neural network, every cube corner should be defined as an isolated region in the second neuronal layer. There are 2^D corners, thus an exponentially large number of isolated regions are needed. The above example illustrates mathematical reasons for the learning disability of neural networks that are based on the concept of gradient learning of discriminating surfaces (including feedforward networks). Other examples as well as formal mathematical proofs of the combinatorial nature of the training requirements for this type neural networks have been discussed in the literature (Minsky and Papert, 1988; Blum and Rivest, 1992).

Now let us determine training requirements in simple cases: How many weights and training examples does a feedforward neural network need for constructing/recognizing a single isolated bounded region in classification space? Delineating a bounded region in a D-dimensional space requires on the order of $\sim 2D$ hyperplanes: two hyperplanes per dimension. For example, a $D = 3$ cube is bounded by six planes. Each $(D-1)$-dimensional hyperplane in a D-dimensional space is defined by D parameters, e.g., by a D-dimensional

vector perpendicular to the hyperplane. Therefore, a feedforward neural network needs at least an order of $2D^2$ weights to define $2D$ hyperplanes for each bounded classification space region. To establish a minimal training requirement for learning an isolated bounded region, note that the required minimal number of training samples is larger than the number of weights, otherwise feedforward neural networks tend to memorize exactly all the training samples and fail to "generalize" or to classify correctly any new sample (Lang et al., 1990). At least an order of magnitude more training samples than the number of weights in a neural network is required for a robust generalization. Thus, even a simple classification geometry of a single bounded region requires $\sim 20 \cdot D^2$ training samples, which is prohibitive for classification spaces of hundreds of dimensions.

It should be emphasized again that the above limitations are fundamental and originate in the limited a priori knowledge utilized by feedforward backpropagation neural networks. These limitations cannot be alleviated by modifications of the learning algorithm that improve convergence, or by using recurrent neural architectures. Using higher order neural networks, calculating terms $\sim W^*_{ii'}x^*_i x_{i'}$ etc. in addition to terms in Eq. (2.3-1), does not alleviate the learning disability either, because the number of such terms is large, on the order of D^2, and grow exponentially with order of nonlinear terms.

Let us summarize. Neural networks that learn discriminating surfaces can be quite efficient in simple cases. And extensive experience may allow for determining the most important terms (neurons in the hidden layer) for a particular problem, thus reducing training requirements. Nominalistic concept of learning implemented by these neural networks can lead to discovering classes in certain simple cases, and thus to formation of concepts. But the structure of these neural networks does not facilitate incorporation of complicated a priori knowledge. And neural networks of this type cannot learn complex classes or concepts on their own. This deficiency is related to the general inadequacy of the nominalistic philosophy, which cannot explain learning of complex concepts by mind.

2.4 Rule-Based Artificial Intelligence

2.4.1 Minsky, Apriority, and Adaptivity

Early neural network and cybernetics research demonstrated adaptive solutions of limited classes of relatively simple problems, but could not be applied to solving the more complicated problems routinely performed by humans. Analyzing their limitations, Minsky came to the conclusion that the development of truly intelligent machines should be based on utilization of extensive a priori knowledge rather than on self-learning. Self-learning does not have a principal role in intellect, suggested Minsky, illustrating his point by an example of Newton's laws—Newton discovered these laws ("self-learned" in this terminology), but our knowledge of these laws is acquired ready-made from textbooks and is an a priori knowledge. Since self-learning is very rarely achievable even by human intellect, solving a self-learning problem by computers, maintained Minsky, was not principally important and was technologically premature.

A framework for incorporation of extensive a priori knowledge into a recognition process was described in Minsky (1968a). Called frame theory, it was based on utilizing a priori information in the form of rules, simulating a process of deliberate rational thinking

by a human. This approach to modeling intelligence got the name "symbolic AI" or "rule-based AI." Throughout this book I call it "rule-based" AI rather than "symbolic" because the nature of symbol is much more complicated than logical rules.[5] In rule systems, categories and concepts are defined a priori, like *Eide* or Ideas in Plato's concept of mind. This new approach proved to hold the key for successful application of computer technology in many areas. Over the following 10 to 20 years, a large number of small- and large-scale AI systems were developed in areas including computer design, engineering design, geophysical exploration, factory automation, military operations, and many others. Rule-based AI research has addressed and solved a number of complicated problems, but some of them turned out to be unsolvable within the rule-based AI paradigm. A number of controversial issues were widely discussed in the scientific literature, often with passionate debates, including proponents and opponents on several sides. Some of these concepts and controversies are discussed in this section.

The main controversy about the rule-based AI concerned the issue of learning and adaptation. Rule systems are most useful when the problem can be exhaustively analyzed during system design and clear rules can be found for every possible case. When uncertainty is an inherent part of the data and unexpected situations are often encountered, rule systems become brittle and inflexible. Minsky's initial attitude toward the problem of adaptation to the ever-changing world and to related stochastic uncertainties was to assume that more and more complex heuristic methods will permit us to get rid of these problems within the framework of rule systems. Reality turned out to be more mischievous than Minsky assumed: uncertainties in data and inaccuracies in solutions, accumulating at each step in the chain of logical inferences built by a rule system, required more and more detailed rules, leading to a combinatorial (or exponential) complexity of the logical inference, which is physically unrealizable.

A rule-based AI approach to utilization of extensive a priori knowledge represented the next attempt after McCulloch to understand intelligence on the basis of realistic philosophy. Similarity between rule systems and Plato's conception of mind based on a priori ideas was discussed by Chomsky (1972). He directly related the principle of apriority in algorithm design to the philosophy of realism. He also hoped that the problem of learning could be solved using a rule-based approach to intelligence. Chomsky attempted to explain language learning based on the theory of language faculty containing a priori linguistic knowledge. The language faculty was understood as a system of grammatical rules (Berwick, 1982).

However, combining grammatical algorithms in a unified system of the language faculty involved the same obstacles that had confronted other adaptive algorithms proposed for rule-based AI: combinatorial computational complexity. The mathematics of rule systems is inadequate for adaptation and learning. This was emphasized by Minsky (1975), the founder of this approach to computational intelligence, and was confirmed in multiple attempts to solve the problem of learning on the basis of rule systems (Winston, 1984). Chomsky came to a similar conclusion, and later he proposed a different approach to the problem of learning based on a priori principles and adaptive parameters (Chomsky, 1981). This principles-and-parameters approach in linguistics was similar to parametric model-based approaches to combining apriority and adaptivity in pattern recognition. It was a step in the direction of model-based learning developed in this book, but, as discussed in the next section, existing mathematical methods in the 1980s used for this purpose were inadequate and faced combinatorial computational complexity.

Chomsky's attempt to base learning on rule systems was one among numerous attempts to add adaptivity to the apriority of the Plato–Minsky rule-based AI. Winston (1984) notes that learning in many algorithms encounters barriers related to the problem of computational complexity: the number of required operations grows combinatorially. Analysis of adaptive algorithms proposed for rule systems, which at first did not seem to be combinatorial, showed this to be an illusion because variabilities and uncertainties were ignored in the original simplified examples. A need to account for uncertainties at the intermediate computational steps turned every algorithm of this kind into a combinatorial one, which is not physically realizable. This was discussed by a number of researchers (Chapman, 1987; Maes, 1991; Franklin, 1995).

Many of the learning systems being developed today still utilize modifications of mathematical methods related to Plato's conception of mind. Looking back at the many years of failed attempts to develop a mathematical theory of learning based on rule systems, it is surprising that the fundamental nature of the difficulties still has not been fully appreciated. The mathematical progress could be faster, if the close relationship between mathematical and philosophical concepts is understood. This is why I would like to summarize again the Aristotelian account of why learning cannot be achieved in Plato's theory of mind. In Plato's theory, there could be no learning, since Ideas (or concepts) are given a priori in their final forms of eternal unchangeable truths. Thus, learning is not needed and is impossible, and the world of ideas is completely separated from the world of experience. This difficulty was "rediscovered" in rule-based AI: new "learned" concepts constructed according to rules do not necessarily correspond to objects in the real world. To remedy the situation, special procedures have to be designed called "symbol grounding." The adaptive relationship between the concepts of mind and objects of the world is treated as an afterthought in rule-based AI as well as in Plato's theory of mind. The intrinsic connection between the world of concepts and the world of objects is missing. It is truly amazing that the impossibility of learning within Plato's theory of intellect was so clearly identified 2300 years ago by Aristotle!

Among the most important achievements of rule systems in the 1970s was that it initiated concerted research on the *internal representation of knowledge*: "many researchers set aside their interest in the study of learning in favor of examining the representation of knowledge in many different contexts and forms. The result was . . . new and powerful ideas—among them frames, conceptual dependency, production systems, word-expert parsers, relational databases, K-lines, scripts, nonmonotonic logic, semantic networks, analogy generators, cooperative processes, and planning procedures" (Minsky and Papert, 1988). Some of these concepts are considered in the following sections.

2.4.2 Soar Production System

Newell and Simon were among the founders of AI. They believed that logic is a fundamental paradigm of human thought. In the 1950s, they developed a computer algorithm, General Problem Solver (GPS), that was capable of proving theorems in logic. They considered computers and the human brain as symbol-manipulating systems, or "physical symbol systems." By symbols they meant abstract mathematical notations such as used in mathematical logic and algebra. Let me repeat again that this designation is fundamentally limited: symbols are much more complex entities than signs used in mathematical notations.

A more general intelligent system evolving Newell and Simon's ideas was developed by Laird in 1981. Its name, SOAR, originally stood for State, Operator, And, Result, its basic problem-solving unit. Since then, it was in continuous development. Currently, "Soar" is a name rather than abbreviation. Soar is intended as a paradigm of general intelligence, a general cognitive architecture for solving problems in any domain. Its view of intelligence is a deliberative logical decision process, whose goal is to solve a specific problem. In a model of the world in Soar, its representation of a problem-solving situation is called a *state*. The world model is what you might expect it to be: for an example of a robot, it may include objects that it perceives around itself. The set of states is called a *problem space*. This is a theoretical notion that does not exist in Soar explicitly. Transits from one state to the next is accomplished by *operators*. A desired state is called a *goal*. Reaching the goal-state constitutes the solution to a problem.

Problem solving in Soar consists in formulating, selecting, and applying operators to a state. This is called a series of decisions. A basic decision cycle involves *productions* and *preferences*. Productions are if–then rules: if C then A. Conditions C refer to characteristics of objects present in the current state, including goals, etc. Actions A propose changes to a state, called *preferences*. First, all productions try to match their conditions, C, to the current state and create their preferences in the preference memory. Some productions may retract preferences. Second, preferences are evaluated and the best (or acceptable to all) action is selected; it is called an operator, and it leads to a change of the state.

Separating productions from operators differentiates Soar from many other production systems. It makes it possible to represent actions without taking them. This enables the internal deliberation during the decision process. An operator may act in the world, or may add an object to a state, which will initiate new productions.

The Soar architecture includes three types of memories and input/output (I/O) (Fig. 2.4-1). Working memory contains a short-term knowledge: the current state including input-perceptions, production matches, output-motor commands, and internal intermediate data structures. Working memory consists of elements, with each element a simple object/identifier-attribute-value relationship (such as i7 ^color red; i7 refers to a particular object). Production memory containing productions is the only long-term memory of Soar. I/O inputs sensory information into the current state and outputs motor commands to the actuators in the world.

A decision process may come to an impasse, for example, when there are no sufficient preferences to select a single operator among several proposed, or when not a single operator is acceptable. In this case Soar creates a subgoal to resolve an impasse. Creation of a subgoal leads to a new set of production matches. If a subgoal is achieved, an impasse is cleared; alternatively a new impasse may appear. This leads to a hierarchy of subgoals. The top goal is generated by a human designer. A hierarchy of subgoals is automatically generated by Soar. A hierarchy of subgoals decomposes a task into subtasks, subsubtasks, and so on. Subgoals may be generated from a set of predetermined subgoals. Alternatively, Soar can use general goal-search methods. An example is a gradient or hill-climbing; it requires a measure of distance from any state to a goal-state as a function of state parameters.

Soar is also capable of learning. Learning is changing the permanent memory, and consists in adding new productions. Soar employs one learning method called *chunking*. It works as follows. When Soar clears an impasse, the chunking mechanism creates a new production, whose condition C corresponds to the state leading to the impasse, and whose

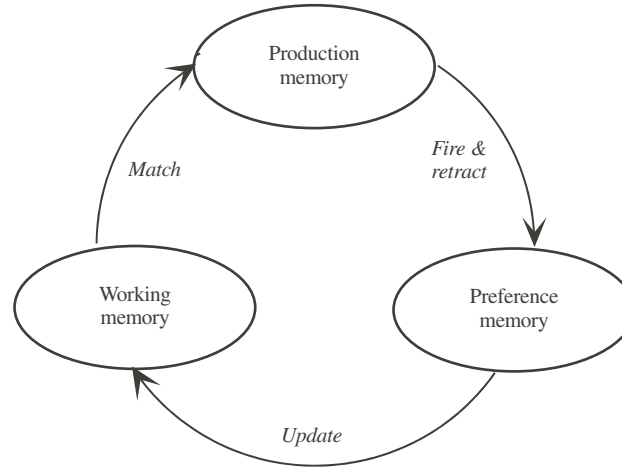


Figure 2.4-1 Soar top-level architecture.

action combines all the actions that have led to the impasse clearing. If the same state occurs again, it does not lead to an impasse, since the new production suggests an adequate action.

As we see, Soar uses a small number of distinct architectural mechanisms. A problem space is a framework for every task. All temporary knowledge is represented in the working memory by objects characterized by attributes with their values. All permanent knowledge is represented by *productions*. All intermediate goals are generated by a single mechanism of automatic subgoaling. And its only learning mechanism is chunking.

According to Soar's authors, they would like Soar to be able to work on the full range of tasks, including complicated open-ended problems; represent and use appropriate forms of knowledge, such as procedural, declarative, episodic, and iconic; employ the full range of the problem-solving methods; interact with the world; and learn about all aspects of the tasks and about its own performance. Important aspects of a general intelligence that are currently missing include the following. Soar does not create its own representations. And its knowledge retrieving is often too slow.

I would add that Soar suffers from general problems of logical rule systems: deliberative thinking is brittle and inflexible; combining it with learning leads to combinatorial complexity. Productions combine specific declarative object-knowledge with specific actions. But they are not adaptive: their conditions are either satisfied or not and they cannot see anything even slightly different from their conditions; multiple conditions lead to a degree of adaptivity, but a very limited one: using multiple conditions to describe a continuum of uncertainties cannot succeed. This leads to a combinatorial proliferation of productions in the presence of uncertainty. These difficulties are of general nature, related to the limitation of Platonian philosophy with respect to learning.

Soar is a continuously changing system. Will its architecture, based on a deliberative logical argumentation, turn out to be too inflexible for the general paradigm of intelligence? Or will it be able to modify to incorporate neural structures and more powerful learning and adaptation methods?

2.5 Concept of Internal Model

The concept of an internal model is among the most important concepts of mathematical intelligence. According to this concept, the entire functioning of an intelligent system (including perception, cognition, prediction, planning, etc.) is based on an internal model (or multiple submodels). Significant aspects of internal models are inborn, a priori. The following main questions should be answered: Which aspects of models are a priori and which are learned? And what is the nature of the a priori representation, so that it supports learning? Various answers to these question relate mathematical internal models to Platonian Ideas, Aristotelian Forms, and Jungian archetypes. Logical rules considered in the previous section are a particular case of internal models related to Platonian Ideas. Here, we overview the historical development of this mathematical concept, including successes and difficulties encountered by various approaches to utilize internal models in intelligent data processing and decision making.

2.5.1 Prolegomena: Parametric vs. Nonparametric Estimation

In the AI community during the 1960s, "intellectual battle lines began to form along such conceptual fronts as parallel (connectionist) vs. serial (symbolic) processing, learning vs. programming, and emergence vs. analytic descriptions" (Minsky and Papert, 1988). At the same time, in several fields of applied mathematics, including financial and economic prediction, pattern recognition, and signal and image processing, conceptually related extensive discussions were carried out concerning relative merits of parametric vs. nonparametric techniques (Tukey, 1960, see 1977). These early discussions revealed methodological differences underlying the two approaches. In parametric techniques, a mathematical (or statistical) model is developed for the problem under analysis based on statistical, geometric, physical, or other phenomenological considerations. A relatively small number of unknown model parameters is estimated from the data. Parametric approaches require a priori analysis and understanding of the problem and can lead to fast real-time adaptation. In nonparametric approaches, no a priori model is postulated, and the large number of estimated parameters is not directly related to the underlying process or phenomenon. Nonparametric approaches were considered suitable for initial, exploratory stages of analysis. Examples of nonparametric techniques in signal processing are Fourier transforms, in which Fourier coefficients are not parameters of any physical model, and parametric autoregressive analysis assuming an autoregressive model of a signal source. In prediction, parametric linear regression is based on a Gaussian model of data variabilities. In classification, linear and quadratic classifiers are parametric techniques based on a Gaussian model of the data variabilities. Examples of nonparametric techniques include those based on the nearest neighbor and discrimination surface concepts discussed in previous sections.

Classical examples of parametric techniques are linear and quadratic classifiers based on the a priori statistical model of Gaussian distributions discussed in Section 1.3. Adaptation or learning consists in estimating parameters of the distributions. The number of these parameters, $N\text{par}$, is relatively small, growing slowly with the problem dimensionality, D: for linear classifiers $N\text{par} \sim D$; for quadratic classifiers $N\text{par} \sim D^2$. Utilization of optimal statistical learning techniques results in fast learning that requires few samples—

the required number of training samples, NT, remains essentially constant independent of dimensionality for the linear classifier, $NT \sim$ constant, and it grows linearly for the quadratic classifier, $NT \sim D$. Using a priori statistical models thus leads to fast and physically acceptable adaptation. These models, however, are too simple for modeling the intellect: classifiers based on these models are limited to simple shapes of decision regions in classification spaces.

The differentiation between *simple* parametric and nonparametric approaches, as discussed, seems to be a methodological one, while fundamental philosophical or mathematical differences are muffled. For example, in signal processing an autoregressive analysis is now often used as an exploratory method, with many parameters that are not necessarily related to an underlying model. From a mathematical standpoint, it seems that a most important difference between the two approaches is how fast the number of parameters grows with the growth in problem complexity. However, later development toward modeling intellect and involving complicated models for large-scale problems revealed the fundamental nature of the differences. The first indication was that training requirements for nonparametric techniques were prone to combinatorial explosion, essentially making them useless for complex problems and as a mathematical apparatus for describing intellect. Model-based techniques could combine apriority of models with adaptivity of model parameters, but early models were too simple for modeling the intellect. We will see in the following sections that utilization of complex models requires development of new mathematical techniques and commitment of considerable resources, and could not have been accomplished without a fundamental shift in the AI paradigm. A shift of the AI paradigm in the 1960s from the connectivist to the rule-based one, from self-learning based on experience to preprogrammed rule systems, was related to the philosophical shift from nominalism to realism. This shift influenced the development of complicated model-based techniques.

2.5.2 Model-Based Vision (MBV)

Success of rule-based AI in the 1970s was extended toward machine vision by developing geometric object models and Model-Based Vision (MBV) recognition techniques (Nevatia and Binford, 1977; Brooks, 1983; Winston, 1984; Grimson and Lozano-Perez, 1984; Chen and Dyer, 1986; Michalski et al., 1986; Lamdan and Wolfson, 1988; Negahdaripour and Jain, 1991; Bonnisone et al., 1991; Segre, 1992; Keshavan et al., 1993; Califano and Mohan, 1994). An MBV paradigm is illustrated in Fig. 2.5-1. It is characterized by an iterative loop of four steps: (1) generating a hypothesis about an image content; then, based on a model corresponding to the hypothesis, (2) predicting features expected in the image and (3) describing features observed in the image in terms of the model components; (4) matching predicted and described features, then, refining the hypothesis and repeating the four steps. Steps 2 and 3 represent an important conceptual advancement in AI: a counterflow of processing is combined, the top-down processing from the models toward the image with the bottom-up processing from an image toward the model elements (features).

Instead of continuously varied parameters of simple model-based techniques considered in the previous section, variability in MBV models is achieved by considering composite models that are composed of various submodels. These high-cognitive level, syntactic models of objects are similar to rule systems. Successes and difficulties of MBV are similar to those of rule systems: MBV techniques do not need training data and do not learn

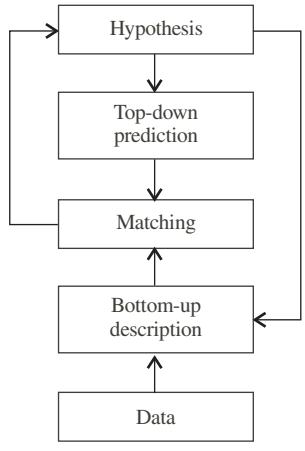


Figure 2.5-1 A model-based recognition concept.

on their own. MBV is successful in structured environments with well-defined objects and constrained variability. When variability increases, it becomes exceedingly difficult to build more and more complicated models and to reason about them. Combining top-down with bottom-up processing in the prediction and description steps helps to mitigate this problem. Another approach to limit real-time computational requirements is to select a limited initial set of object hypotheses using bottom-up processing techniques, which is called indexing (Lamdan and Wolfson, 1988; Califano and Mohan, 1994). But when unforeseen variabilities are a constant factor in the recognition problem, MBV faces difficulties that are common to rule systems. More and more detailed models are required, potentially leading to a combinatorial explosion of model complexity. MBV models in their completeness resemble Plato's ideas, and mathematically are similar to rule systems. Thus, the MBV method faces difficulties that are generic to the Plato–Minsky rule-based paradigm of intellect, which treats adaptation as an afterthought.

2.5.3 Adaptivity and MBV

Whereas simple parametric models emulated adaptivity of intellect in simple situations, complicated rule-based models discussed above emulated the intellect's apriority. To combine their advantages and overcome their limitations, parametric model-based techniques have been proposed to combine the adaptivity of parameters with apriority of models. In this approach, adaptive parameters are used to adapt models to variabilities and uncertainties in data. A basic mathematical apparatus for achieving this synthesis is called the Multiple Hypothesis Testing (MHT) algorithm formulated in Section 1.3.3.

MHT attempts to combine multiple adaptive parametric models into a complex world model. In the process of learning, the MHT algorithm selects the most appropriate model for each piece of data, which is achieved by checking all (or a significant part of) possible associations or combinations of models and data. By utilizing complex adaptive models, MHT combines both factors of the intellect, the apriority of models and adaptivity of model parameters. However, MHT leads to another type of limitation: computational complexity. The number of required elementary operations in the MHT algorithm grows

combinatorially as a function of model complexity and becomes too large even for relatively simple problems. For realistic problems of medium complexity the number of required computations becomes *unphysically* large, exceeding the number of all interactions between all elementary particles in the entire history of the universe[6]; for this reason the MHT method is also not suitable as a foundation for the physical intuition of the nature of mind. Let us repeat again, this limitation of the approach based on models specified in their final crisp forms is related to the general limitation of the Platonian conception of mind based on ready-made Ideas that are given a priori, before any experience.

2.6 Abductive Reasoning

2.6.1 Deduction, Induction, and Abduction

Let us overview the three common ways of reasoning. Deduction is reasoning from general to concrete. For example, let A be a statement defining an apple: "an apple is red, round, and sweet." And let X be a scene. The deductive reasoning goes as follows: if X contains an apple, then X should contain a red and round object, and this object should be sweet. Deduction is a top-down reasoning. Deductive approach to designing intelligent systems suggests that a robot should be supplied with a priori knowledge (a database) of general concepts. This is the approach of expert systems and rule-based AI. It works well when there is little variability in the data, and every object is exactly as described by the general concept. When there is variability or uncertainty, the deductive approach is prone to a combinatorial explosion of the complexity and number of rules. This approach is related to Plato's philosophy of realism: general concepts are given to the mind a priori, before any experience.

Induction is reasoning from the concrete to the general. In our example of apples, induction goes as follows. I see one red round object, I tasted it, and it is sweet. I see another red round object, I tasted it and it is sweet. I may conclude that every red round object is sweet, that they make up "a class of objects," A, that I will arbitrarily call, say "yabloko" (this means apple in Russian). An inductive approach to designing intelligent systems suggests that a robot does not need any a priori knowledge of general concepts. Robot "just sees" similarities, and it groups (or clusters, or classifies) objects according to their similarities. Then some arbitrary names are assigned to these classes (in order to understand each other, of course, robots, like people, will have to agree on some arbitrary convention of names). This is the approach of self-learning clustering algorithms, such as the nearest neighbor. It works well when there are few clearly distinct types of objects, which can be differentiated using few distinctly perceived attributes. When there are a lot of attributes, and objects may be similar in some and dissimilar in other attributes, the inductive approach faces combinatorial explosion of the complexity of learning and of the number of required examples to learn. This approach is related to Antisthenes' philosophy of nominalism: general concepts are names, arbitrarily given to classes of objects, and the classes are groups of objects that look similar.

Abductive reasoning about our apples goes as follows. Given the general definition A, and a red round object, I may conclude that this object is an apple. Note the difference in the direction of inference comparatively to the deduction: here the inference is from

the concrete to the general. But, unlike induction, it is according to the general law, which should be given a priori. Abductive reasoning is also called reasoning by analogy: I conclude about the general concept, which governs the concrete case, by analogy with other cases, where this general concept has been applicable in the past. Some people consider abductive reasoning to be the only way of creative reasoning: new general concepts can be created by analogy with the old ones, by differentiating the old concept. For example, one can perceive the similarity among red round objects according to the general concept, and then learn differences between apples and nectarines from experience. Thus a concept of "nectarine" is learned.

2.6.2 Abductive Reasoning Trees and Bayesian Networks

Abductive reasoning combines apriority of models with adaptivity to the data, and in this way it is an adaptive model-based reasoning. Three types of models are in popular use for abductive reasoning: logic-based models, probability-based modes, and a combination of the two. A typical application for abductive reasoning is to find structural relationships among variables in a database, given certain a priori knowledge of this structure. The variables and relationships are the ingredients of the model. The relationships are the domain's knowledge, which should be improved using the existing data. An additional aspect of the domain knowledge is that of uncertainty about the relationships and conditions of their applicability. Logic-based models are good for representing qualitative structural knowledge, but could be cumbersome in handling uncertainties. Probability-based models are good for representing uncertainty, but do not represent the qualitative structures as well. A method for combining these advantages was developed by Bhatnagar and Kanal (1993).

The Bayesian equation for the a posteriori probabilities that we discussed in Chapter 1 (1.2-15) is a classical example of abductive reasoning, from the data to the models, while accounting for the restrictions imposed by the models. Since the Bayesian equation is a consequence of the rule of conditional probabilities, it is very convenient to represent the a priori knowledge about the domain structure in terms of conditional probabilistic relationships. A joint probability (or probability density) of n variables can be written as follows:

$$P(x_1, x_2, x_3, \ldots, x_N) = P(x_1|x_2, x_3, \ldots, x_N)^* P(x_2|x_3, \ldots, x_N)^* \cdots^* P(x_N) \quad (2.6\text{-}1)$$

This equation is obtained from the rule of conditional probability (1.2-2) by applying it sequentially, as if in a "chain," and it is called the chain rule. It is always valid; there is no limiting assumption or specific domain knowledge incorporated in this relationship. Learning or estimating this probabilistic model from the data will involve a combinatorial explosion of training requirements characteristic of classical pattern recognition algorithms. This is why the model-based approach consists in using restrictive models, which already contain a significant amount of information about the domain. For example, if there are reasons to believe that in a particular application domain all the relationships among variables are pairwise, an appropriate model is given by

$$P(x_1, x_2, x_3, \ldots, x_N) = P(x_1|x_{m1})^* P(x_2|x_{m2})^* \cdots^* P(x_N) \quad (2.6\text{-}2)$$

Here, general conditional probabilities are replaced by pairwise relationships; it is quite a restrictive general type limitation on the structure. And it contains only $(N - 1)$

relationships, $P(x_n|x_m)$. Usually, this type of model is further restricted by requiring that each variable appears only once in the n-position, and after it does (in the left-to-right direction along the chain), it is not used anymore in the m-position. With this restriction, the model (2.6-2) is a "tree," as illustrated in Fig. 2.6-1.

A more general type model can be written as

$$P\,(x_1, x_2, x_3, \ldots, x_N) = P\,(x_1|S_1)^*\,P\,(x_2|S_2)^*\cdots^*\,P\,(x_N) \tag{2.6-3}$$

Here, S_m is a subset of all variables $\{x_{n'}\}$, on which variable x_n directly depends. This more general model is called a Bayesian network (there are similar restrictions on variables in n- and m-positions).[7] A Bayesian network represents the domain knowledge of structural relationships in terms of probabilistic conditional dependencies and independencies. The probabilistic interpretation facilitates consistent computation of probabilities from the data.

Having described structural models that are typically used in abductive reasoning, let us consider a problem of inferring "concepts" underlying the structure of data from a database. Or, in other words, determining the detailed structure from the data, based on some a priori assumption about the structure. This is approached by designing a certain criterion, which should be maximized in the process of learning the structure. Such a criterion to some extent depends on what is desired to be achieved. For example, Bhatnagar and Kanal (1993), wanted to achieve an approximation of the distinct qualitative relationships, so that all $P\,(x_n|S_m)$ are as much as possible close to either 1 or 0. If each data base case (x_n) corresponds to a separate "explanation" S_m, then the probabilities can be defined to be 0 or 1, but in order to have an explanatory power, a single "explanation" S_m must be suitable for more than one case. Their procedure for achieving a consistent estimation, in a somewhat simplified way, is to minimize the average conditional entropy

$$E = -\sum_{n;m} P\,(x_n|S_m)\,\ln\left[P\,(x_n|S_m)\right] \tag{2.6-4}$$

Minimization of this criterion can be interpreted as minimization of the uncertainty with which $x_n$follows from S_m. The algorithm used for the minimization of the entropy and

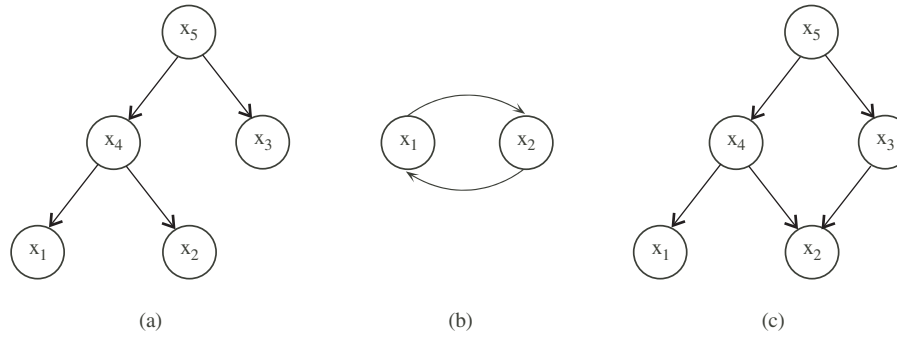


Figure 2.6-1 Illustration of trees and Bayesian network structures. (**a**) A tree and a Bayesian network; (**b**) not a tree, not a Bayesian network; (**c**) a Bayesian network, but not a tree.

estimation of the model parameters (sets S_m) is described in Bhatnagar and Kanal (1993). In this algorithm, learning is combinatorial, as in other model-based approaches described in Section 2.5 due to the need to consider combinations of subsets of $\{x_n\}$ that make up sets S_m.

Bayesian networks can be used as a consistent modeling tool. The technique described above can be viewed as a model-based clustering, where a Bayesian network is used as a model. Clusters are formed by those x_n, which have high a probability of being "explained" by the same model S_m. Model-based clustering therefore can be called adaptive model-based reasoning. Conversely, an adaptive model-based reasoning is a form of abductive reasoning: it combines knowledge with learning in that concepts-clusters are formed using a priori models, while parameters of the models are estimated adaptively from the data. Bayesian network models possess a desirable feature, they can naturally be used to form hierarchies: every subset S_m can be composed of hierarchically restricted subsets. A simple case of a hierarchy is a tree model. The importance of the hierarchical models for building complicated intelligent systems will be discussed in Section 2.12. The technique described above, however, led to combinatorial complexity. Although a step toward the Aristotelian theory of mind, it relies on specific crisp structures, and in that is similar to the Platonian conception of mind and to model-based approaches considered in the previous section. Modeling field theory, described in Chapters 4 through 8, can be viewed as a model-based clustering, in which combinatorial search is eliminated. A challenge remains to develop modeling field theory for hierarchical internal models, like Bayesian networks, or trees in such a way that hierarchies will emerge in the process of learning without combinatorial searches.

2.7 Statistical Learning Theory and Support Vector Machines

2.7.1 Model Complexity: Risk Minimization vs. PDF Estimation

Estimation of probabilistic models or pdf of data is a general approach to solving a wide range of problems in the areas of classification, recognition, prediction, and data mining. Optimal solutions to these problems can be devised in terms of pdf. Estimating pdf in high-dimensional spaces, however, involves the difficulty of combinatorial complexity: estimating pdf in a general form without any constraints involves the "curse of dimensionality"; imposing model-based constraints with flexible models composed of multiple submodels leads to combinatorial complexity of computations. Another approach to limiting complexity of the estimation process, called Statistical Learning Theory, was developed by Vapnik and his co-workers. It is also known as the Vapnik–Chervonenkis (VC) theory. Vapnik considers pdf estimation an overkill: pdf estimation might take more training samples than designing a classifier or predictor. Therefore, he considers instead a problem of risk minimization, where risk is an appropriate measure for a specific application. For example, the maximum likelihood estimation can be considered as risk minimization if risk is defined as negative likelihood. This section follows Vapnik (1995) and Cherkassky and Mulier (1998). VC theory is often referred to as complicated and difficult to understand, therefore here I am concentrating on the main concepts and results, which are quite clear and intuitively obvious.

A general formulation of a risk-minimization problem is as follows. We would like to predict y given $\mathbf{x}$. For example, y could be a Dow Jones value (continuous) or a class of objects (discrete number) and x could be the past values of Dow Jones, interest rates, etc. During training, we have a set of training data $\mathbf{Z}_N = \{(\mathbf{x}_n, y_n), n = 1, \ldots, N\} = \{\mathbf{z}_n\}$. Predictions are made by prediction function, $M(\mathbf{z}, p)$ that depends on parameters, p, which should be estimated during training or learning process from training data. (Parameters p generally is a multicomponent vector, but for simplicity of notations we do not use bold for p; the same is true about y.) For the example of Dow Jones prediction, an appropriate risk measure $R(p)$ is the expected error,

$$R(p) = E\left\{[y - M(\mathbf{z}, p)]^2\right\} \tag{2.7-1}$$

Considered prediction functions belong to a set of prediction functions, which could include all possible functions, or could be limited, say, to a specific functional shape. It is convenient to perform the analysis by using loss functions, $q(\mathbf{z}, p)$, which depend on the prediction function and on the type of the risk that is considered. For the above example,

$$q(\mathbf{z}, p) = [y - M(\mathbf{z}, p)]^2 \quad \text{and} \quad R(p) = E\{q(\mathbf{z}, p)\} \tag{2.7-2}$$

The loss function depends on the parameter values p; as p values vary, the loss functions vary over the set of loss functions, $Q = \{q(\mathbf{z}, p)\}$. The final objective is to select parameter values p that minimize the loss (2.7-1), however, the loss measure is unknown, because it is an excepted value over the unknown pdf($\mathbf{z}$). Therefore, practically, we are limited to using empirical risk, the risk averaged over the available training data,

$$R_{\text{emp}}(p) = (1/N)\sum_n q(\mathbf{z}_n, p) \tag{2.7-3}$$

Vapnik calls minimization of (2.7-3) empirical risk minimization (ERM). In general, the minimal empirical risk is too optimistic, $\min R_{\text{emp}}(p) < \min R(p)$. This is because it is minimized over a limited number of samples and parameters p tend to be "overfitted" to a particular training set. For example, with the nearest neighbor type training, it is easy to achieve $\min R_{\text{emp}}(p) = 0$. Of course, this perfect performance would not generalize to a new data point. The very first theoretical problem, therefore, is to establish conditions under which ERM leads to an approximate minimization of the true risk (2.7-2). This requirement can be formulated as follows. Let p^* be the parameter values minimizing the empirical risk. It is desirable that as the number of samples grow, the empirical risk and the true risk converge to the true minimal risk value $R(p_o)$ (over the given set of loss functions Q):

$$R(p^*, N) \to R(p_o) \text{ and } R_{\text{emp}}(p^*, N) \to R(p_o), \quad \text{as } N \to \infty \tag{2.7-4}$$

This requirement is called consistency of the ERM. Statistical learning theory established conditions of consistency of the ERM and bounds on its generalization ability, developed inductive principles of learning from small training samples consistent with these bounds, and developed constructive methods for implementing these principles. I will briefly review some of these results.

2.7.2 Consistency of ERM and VC Dimension

A nearest neighbor training example, as mentioned above (and in Problem 2.7-1), leads to the perfect performance on the training data set, but does not generalize to new data. The problem is that in the nearest neighbor training, the resulting set of loss functions is too flexible: it can perfectly fit any training dataset, and, therefore, from the training data we never now if the true risk minimization problem is easy or hard. An important idea of the statistical learning theory, which it shares with all adaptive model-based learning approaches, is that the flexibility (or diversity) of the set of models should be matched to the training data set. If the flexibility of loss functions is large relative to the training data, empirical risk will be much lower than the true one. And if the flexibility of loss functions is small relative to the training data, empirical risk will be close to the true one; however, the true minimal risk achievable with the given set of functions will be too high relative to what could be achieved with more flexible functions.

To be able to match flexibility of loss functions to the size of the training data set, one needs to have an independent measure of the flexibility of a set of functions. Such a measure, called the Vapnik–Chervonenkis (VC) dimension, is a key result of the statistical learning theory. Let us describe this quantity. A function $q(\mathbf{z}, p)$ can be used to partition a training data set $\mathbf{Z}_N$ into two sets as follows,

$$\begin{aligned} &\text{if } q(\mathbf{z}_n, p) - \text{th} \geq 0, \mathbf{z}_n \in \text{set } 1 \\ &\text{if } q(\mathbf{z}_n, p) - \text{th} < 0, \mathbf{z}_n \in \text{set } 2 \end{aligned} \tag{2.7-5}$$

We call such a partition a dichotomy (in order to attribute the dichotomy to q, we assume that the threshold, th, is included into the set of parameters p). A flexibility or diversity of the set of functions $Q(\mathbf{z}) = \{q(\mathbf{z}_n, p)\}$ can be measured by the total number of different dichotomies $N(Q, \mathbf{Z}_N)$ that a given set $\mathbf{Z}_N$ can be split into by the given set of functions Q.

A growth function $G(N, Q)$ is defined as a logarithm of the maximum number of dichotomies that a given set Q can induce on any set of the size N,

$$G(N, Q) = \ln \max_{\mathbf{Z}_N} N(Q, \mathbf{Z}_N) \tag{2.7-6}$$

What is the maximal value of G for any data set Q? A set of N points can be divided into two subsets in 2^N different ways, thus,

$$G(N, Q) \leq \ln 2^N = N \ln 2 \tag{2.7-7}$$

The necessary and sufficient condition of the consistency of the empirical risk minimization can be formulated in terms of the growth function as follows:

$$[G(N, Q)/N] \to 0, \quad \text{as } N \to \infty \tag{2.7-8}$$

Note that this condition is independent of any specific pdf or training data set and refers only to the size of the training data set. In addition to guaranteeing the consistency of the estimation (2.7-4), the above condition also guarantees the fast rate of convergence defined as follows:

$$P\left[R_{\text{emp}}\left(p^*, N\right) - R\left(p^*, N\right) > \varepsilon\right] \leq \exp\left(-cN\varepsilon^2\right), \quad c > 0 \tag{2.7-9}$$

This reads as follows: the probability that the difference between the true and empirical risk exceeds some small positive value ε goes to zero exponentially as the number of training samples goes to infinity. Because empirical risk is a random number, a small probability cannot be excluded that a large deviation may occur between the empirical and true risk. A higher precision (small error ε) requires a larger number of samples, $N > 1/(c\varepsilon^2)$, and c is a constant. This could be written as $\varepsilon > 1/\sqrt{N}$. Note, a very simple analog of this relationship: expected error goes down as a square root of the number of samples (this relationship is well known for simple cases, such as estimation of the mean by taking an average value). For practical applications, the value of c is important. Vapnik indicated that $c \leq 4$, however, $c = 4$ is a worst case (such as discontinuous pdf); for practical purpose it can be taken as $c \cong 1$ (Cherkassky and Mulier, 1998).

The VC dimension of the set of functions Q is defined as follows. If N samples can be partitioned by Q in all possible 2^N ways, it is said to be shattered by Q. Set Q has a VC dimension h if there exist a set of h samples that can be shattered by Q, but there is no set of $h + 1$ samples shattered by Q. Growth function is related to VC dimension of the set of functions. It turns out that $G(N)$ is limited either as

$$G(N) \leq N \ln 2, \text{ or} \tag{2.7-10}$$

$$G(N) \leq h[1 + \ln(N/h)] \tag{2.7-11}$$

If the VC dimension is finite, than at large N the last condition holds true. Finite VC dimension leads to (2.7-8) and is a necessary and sufficient condition for consistency and fast convergence of the ERM. In addition, the following bound on the true risk was obtained:

$$R(p) \leq R_{\text{emp}}(p) + \Phi(h) \tag{2.7-12}$$

Here Φ is called the confidence interval; it is a growing function of the VC dimension. It follows that the optimal minimization of risk can be formulated as a tradeoff between empirical risk minimization (which is reduced as h increases) and confidence interval minimization (which increases with h).

Examples. The VC dimension is difficult to compute for complicated sets of functions. Table 2.7-1 contains several examples for which the VC dimension is known.

2.7.3 Support Vector Machines (SVM)

Conventional statistical and neural network methods control model complexity by using a small number of features (the problem dimensionality or the number of hidden units). SVM controls the model complexity by controlling the VC dimension of its models. This method is independent of dimensionality and can utilize spaces of very large (infinite) dimensions. Utilization of very large dimensional spaces permits constructing a very large number of nonlinear features and then performing "adaptive feature selection" during training. By "shifting" all nonlinearity into the features, SVM can use linear models, for which the VC dimension is known. Only a sketchy outline of SVM is presented here; for more details the reader is referred to Vapnik (1995) and Cherkassky and Mulier (1998).

TABLE 2.7-1
Examples of VC Dimensions for Sets of Functions

Number	Function Type	Math. Definition	VC Dimension
1	Linear functions in D dimensions	$q(\mathbf{z}, p) = \sum_{i=1}^{D} p_i^* z_i + p_o$	$h = D + 1$ (the number of parameters)
2	Rectangular indicator functions in D dimensions	$q(\mathbf{z}, c, w) = 1$, if $c_i - z_i \le w_i$ $q(z, c, w) = 0$, otherwise	$h = 2 \cdot D$ (the number of parameters)
3	Radially symmetric indicator functions in D dimensions	$q(\mathbf{z}, c, r) = 1$, if $\|\mathbf{c} - \mathbf{z}\| \le r$ $q(\mathbf{z}, c, r) = 0$, otherwise	$h = D + 1$ (the number of parameters)
4	Local functions in D dimensions (G is a Gaussian or any other "local" function)	$q(\mathbf{z}, c, r) = G(\|\mathbf{c} - \mathbf{z}\|/r) - \text{th}$	$h = D + 1$ (less then the number of parameters, $\mathbf{c}, r, \text{th} : D + 2$)
5	Linear combination of a fixed set of basis functions in D dimensions	$q(\mathbf{z}, p) = \sum_{i=1}^{M} p_i^* g_i(\mathbf{z}) + p_o$	$h = M + 1$ (the number of parameters)
6	Linear combination of an adaptive set of functions in D dimensions, $\mathbf{v}$ is a set of adaptive parameters	$q(\mathbf{z}, p) = \sum_{i=1}^{M} p_i^* g_i(\mathbf{z}, \mathbf{v}) + p_o$	Unknown, even if the VC dimension for each g_i is known

Consider a classification problem with two classes; the data are features $\mathbf{x}$ that are vectors in D-dimensional space and class labels y with values ± 1. A linear boundary (a hyperplane) in $\mathbf{x}$-space is given by $(\mathbf{px}) + p_o = 0$. Training data are $(\mathbf{x}_n, y_n)$, $n = 1, \ldots, N$. Here we consider a case of linearly separable data (this restriction is not really required, but this case is easier to analyze). For linearly separable data, class $1 (y_n = +1)$ data points are on one side of the separating hyperplane, $(\mathbf{px}_n) + p_o \ge 1$, and class $2 (y_n = -1)$ data points are on the other side of the separating hyperplane, $(\mathbf{px}_n) + p_o \le -1$. These equations can be written in a compact form,

$$y_n \cdot \left[\left(\mathbf{px}_n \right) + p_o \right] \ge 1 \tag{2.7-13}$$

Let us consider separating hyperplanes satisfying the constraint $\|\mathbf{p}\|^2 \le c$. Here we define the norm as $\|\mathbf{p}\|^2 = \sum_{i=0}^{D} p_i^2$. The VC dimension for separating hyperplanes satisfying this constraint is given by

$$h \le \min \left(r^2 c, D \right) + 1, \quad \text{for large } D, h \le r^2 c + 1 \tag{2.7-14}$$

Here, r is the minimal radius of a sphere that contains all the training data $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$. For large dimensionality D, the VC dimension is independent from D and depends only on $r^2 c$.

Let us analyze the meaning of this expression for the VC dimension. Note a simple geometric fact: a distance from a point x' to the hyperplane $(\mathbf{px}) + p_o = 0$ is given by $[(\mathbf{px}') + p_o]/\|\mathbf{p}\|^2$ (see Problem 2.7-2). Parameters p are measured in units of inverse x, therefore, $(1/\|\mathbf{p}\|)$ has a unit of length in x-space. Although there is no intrinsic length associated with the hyperplane $(\mathbf{px}) + p_o = 0$ (scaling p by any value does not change this equation), there is an intrinsic length associated with Eq. (2.7-13). Training data points

that turn this inequality into an equality $[(\mathbf{p}\mathbf{x}_n) + p_o = \pm 1]$ are lying at a distance $(1/\|\mathbf{p}\|)$ from the separating hyperplane. These are the training data points that are most close to the separating hyperplane. They are called the support vectors. Let us denote the distance from the support vectors to the separating hyperplane r_o (instead of $1/\sqrt{c}$ used above). So, the VC dimension of the separating hyperplane satisfies

$$h \leq (r/r_o)^2 + 1 \tag{2.7-15}$$

This is a very intuitive result: it says that the flexibility (or diversity) of the sets of separating hyperplanes is a simple function of the ratio of the distance separating classes, $2r_o$, to the entire extent of the data, $2r$ (in the direction perpendicular to the separation hyperplane). This is an exact mathematical formulation of the notion of simple and complex problems discussed in Section 2.2 (Fig. 2.2-2). The problem is "simple" if classes are well separable (r/r_o is a small number); then the problem can be solved with a low-VC dimension hyperplane. Remember that according to (2.7-12), the optimal solution is a tradeoff between the empirical risk and confidence interval. The empirical risk of a separating hyperplane is 0, therefore we need to minimize the confidence interval, and, correspondingly, we need to minimize the VC dimension. This, in turn, requires us to find a hyperplane leading to the widest possible separation between classes. Such a hyperplane is called the optimal separating hyperplane. For the optimal separating hyperplane, the following bound on the classification error was obtained by Vapnik:

$$E_N\{\text{error rate}\} \leq E_N\{\text{number of support vectors}\}/N \tag{2.7-16}$$

Here, expectations are taken over all training data sets of size N. Again, this is a very intuitive bound: the expected error is proportional to the relative number of the samples that are closest to the separating hyperplane. Note that this expression does not imply that the number of support vectors should be minimized, just the opposite, as argued above, the number of support vectors should be maximized (because we want the widest separability between classes); the above bound is obtained for the optimal separating hyperplane, which is "supported" by the maximal number of support vectors.

Finding an optimal hyperplane requires solving a quadratic optimization problem, which could be a computationally involved task in the spaces of millions of dimensions. This problem was reduced by Vapnik to manageable complexity by "folding" a large number of dimensions into a smaller dimensional space with nonlinear metrics. These algorithms, however, are beyond the scope of this book.

Let us summarize the main points of the statistical learning theory (SLT) and its relationship to the main topic of this book, modeling field theory. SLT provides the mathematical apparatus for learning complicated nonlinear relationships from data, while controlling combinatorial explosion and ensuring the generalization capability of the solution beyond just the training data set. Its historical development was along the lines of "shrugging off" the a priori information and concentrating on learning from the data—which is contrary to the main line of modeling field theory: combining apriority with adaptivity. However, SLT's recent results indicate that a priori information can be quite useful for designing nonlinear metrics that are used to "fold-in" the very large dimensional spaces into nonlinear spaces of smaller dimensions. The original philosophy underlying the development of VC

theory is that of nominalism: learning from data without a priori knowledge. Yet I prefer to consider this theory as a general method applicable to model-based recognition and to determining the optimal complexity of adaptive models relative to the amount of available training data.

2.8 AI Debates Past and Future

This Section summarizes the historical development of the mathematics of intelligence as well as emergent trends in terms of the main lines of arguments, disagreements, and debates. Its position in the middle of the chapter, half-way between the classical and newly emerging concepts, is intended as a review of the classical concepts, their significance for the general problem of the science of intellect, and the difficulties they faced. It also outlines the directions of newly emerging concepts designed to counter the main difficulty of the conundrum of combinatorial complexity, and thus prepares a background for the following sections.

2.8.1 Arguments and Disagreements: An Overview

The beginning of the research field named artificial intelligence (AI) is usually dated to the 1956 summer institute at Dartmouth College. The leading figures in the mathematics of intelligence during the previous decades included Wiener, Neumann, Shannon, McCulloch, and Turing. A new generation that came to lead the field after the 1956 meeting included McCarthy, Minsky, Newell, and Simon. Although every new field of spiritual endeavor has its disagreements, tensions, and detractors, the battles about AI "have been particularly vehement, because of the dramatic promise of the thinking machine" (Minsky and Papert, 1988). The field at the center of inquiry was of the utmost importance and interest for humankind, and researchers from various fields and of various persuasions were eager to stake their claims. Before turning to specific contentious issues in the following subsections, I first present a brief overview.

The main lines of disagreements included the fundamental paradigm of intelligence: some researchers saw this as self-learning applicable to any problem, and others believed in expert machines supplied with detailed knowledge about specific domains. This is the same schism between adaptivity and apriority that we have already discussed. In the 1950s, the approach was to develop self-learning automata that could learn everything from a few basic principles. Toward the late 1960s, the notion of expert systems, machines supplied with a large body of expert knowledge about particular domain, became widely accepted. During the 1980s this trend was reversed with advancements in neural networks. A widespread realization that both aspects, the apriority and adaptivity, are needed came only in the 1990s.

Another battle line was formed along logic vs. neural. The interests of a new generation of AI researchers in the 1950s got gradually tilted toward utilization of logic as a fundamental paradigm of human thought. The view considering intelligence equivalent to logic evolved from Aristotelian logic over 2000 years. At the beginning of the nineteenth century, Boole published a book, *The Laws of Thought,* in which he formalized Aristotelian logic and freed

it from the ambiguities of natural language.[8] Based on his work, Whitehead and Russell in *Principia Mathematica* (1910–13) attempted to derive the entire mathematics from the laws of logic. In 1931, Gödel proved the mathematical futility of this attempt. Nevertheless, the practical consequences of Gödel's theory, which we will discuss in Chapter 11, were not clear to many at the beginning of the AI era. In the 1950s, Newell and Simon developed computer algorithms that proved many of the theorems from *Principia Mathematica*, and logic was widely believed to be the principal paradigm of intelligence.

In the 1980s, new powerful neural network algorithms were developed and the debate was reinvigorated. The debate of neural vs. logic got entangled with the debate on apriority vs. adaptivity. The 1990s saw the emergence of hybrid systems combining both approaches, logical and neural. It is useful to remember that our brain evolved from a number of subsystems, which possibly changed their functions in the course of evolution. And the term hybrid might be quite appropriate. Still to me, the concept of a "hybrid" system is uninviting: it emphasizes the luck of our understanding of how our mind combines these two aspects. Especially so, because we still do not understand the principles of the subsystems that are supposed to be combined in a hybrid. The evolution of the brain was a slow process, and its various modules evolved to coordinate their mechanisms and functions in an amazingly tuned way. In Section 2.14 we will return to this discussion and will emphasize the existence of unifying organizational principles applicable to the diverse brain modules.

Can a machine think? was among the main issues vigorously debated since the 1950s. Proponents of machine thinking included researchers too eager to exaggerate the state of the art and boosting unrealistic expectations, before an in-depth understanding of the complexities of the thinking process was attained. The opponents included people who believed that thinking is a mysterious property of living matter and a machine would never be able to think the way human do. To the opponents, I would offer a question: why are you so sure that humans possess this mysterious property? Since Freud, psychology revealed that our introspection can severely mislead us. I believe very strongly in the power of intuition and introspection as far as they lead to rational explanation. But, as a support for beliefs in mysterious powers, these are very shaky grounds. The resolution of this debate I see in the rational analysis of the thinking process, in the differentiation of "thinking" into constituent components and processes. To the extent that we succeed, we will be able to build thinking machines. To the extent that some aspects of human thinking will remain beyond our understanding, we will have to analyze reasons why this is so, rather than sweep the annoying questions under the rug (but this is so difficult to avoid!). And a step in this direction will be to identify and isolate mysterious aspects from the others that are easier to understand. The debate on "can a machine think?" helps in this analytic process.

Most contentious issues include "representation." Are there internal representations of the world in the mind? Although sometimes debates help clarify the issue, in this case just the opposite happened. Many immensely respectful scientists, especially of the neural bent, would not admit the existence of internal representations. And some staked their careers on disproving anything of this sort. This seems to be their reaction to the previous claims of logical and rule-based AI researchers that the internal representations are the logical rules hardwired in our brains. The point of view in this book is that the internal representations are in the forms of models, which in rare cases could be similar to logical rules, but most often are fuzzy, uncertain, fleeting, and capable of adaptation. Because of the contentious nature of this issue, let me emphasize that internal representations discussed in this book are a

ubiquitous property of the living beings even in the simplest forms. For example, an *E. coli* bacteria sensing a gradient of sugar along its body possesses an internal representation of the world. The gradient is the *E. coli*'s world model, even if it is represented just by a simple change of some chemical inside the *E. coli*. The debate about internal representation started in the 1980s, intensified in the 1990s, and likely will continue for a while.

2.8.2 Can a Machine Think?

I was a kid in a beautiful Odessa city, a port and resort on the Black Sea in southern Ukraine. Odessa maps were state secrets and were not sold in stores. I knew how to get around the city by using trolleys, but I vividly remember struggling in my mind when I tried to visualize the entire city map. I tried to do this long before I had enough knowledge to succeed. Undoubtedly, the human mind possesses an ability to figure out the map from experience. And this ability is based on an inborn (a priori) ability to conceptualize an ordered space (or spatial representation). Kant numbered this ability among the few main a priori concepts.

The struggle in my mind was about how to reconcile two different internal representations of the knowledge about the city. One representation listed trolley stops and the rules of how to get "there" from "here." Another representation was a spatial map. These two representations are essentially about the same knowledge. In addition, the map representation utilizes a concept of an ordered space, and the concepts of distance or similarity, leading to a possibility for fuzzy reasoning. Both representations lead to an understanding, but one understanding is much more powerful than the other.

This example illustrates that an understanding is a matter of degree and that different representations underlie different understandings. There are types of understanding that lead to combinatorial type algorithms, and other ones that lead to efficient algorithms. Remember, driving a car according to instructions ("soon after T.J. Maxx turn left, but if you see a flashing light, you missed your turn, and you should go back"). A set of instructions very quickly grows combinatorially, and still you may get lost. At this moment, everyone would appreciate having a map and enjoy a power of efficient representation of knowledge. This is not a joke anymore: in the presence of uncertainty, logical rules lead to combinatorial complexity. The rule-based AI was broken into pieces over this "simple" point.

One of the most dramatic examples questioning the nature of understanding is due to the philosopher Searle (1980). In his famous "Chinese Room," Searle imagines that he is sitting in a closed room and he is supposed to answer written inquiries in Chinese. He has no knowledge of Chinese, but he has a stack of file cabinets filled with look-up tables, which contain all possible inquiries and answers. For every inquiry, he finds a matching entry in a table and copies the answer. Nobody around can see that he does not understand Chinese. But this kind of knowledge stored in look-up tables is not what we call human understanding. Searle's intention was to prove that computers cannot think the way human do.

This example is not only dramatic, but also preposterous. The preposterous aspect is in assuming that the entire language can be represented as a set of look-up tables, or that any representation, significantly different from the one we actually use, will work at all. The serious and nontrivial aspect of Searle's example is in focusing our attention on what is *understanding*. Philosophers from Aristotle to Kant answered this question

by referring to internal representations: understanding is a set of concepts in our mind along with interrelationships among them. The type of interrelationships is given by the representations.

Several adamant arguments against computer's thinking are discussed by Penrose (1994). For example, he considers a chess position requiring a global analysis, which is obvious to a human and very difficult for Deep Thought (a computer programmed to play chess). In this case, it is obviously a matter of representation. It is easy (in principle) to program a computer so that it would be capable of a specific type of global analysis required by supplying it with an appropriate global representation. But there is no need for these complicated examples: there are more convincing examples closer to home. Elementary arithmetic is made easy by the specific representation of numbers we are using; it is called Arabic numerals, or positional system. Multiplications and divisions of numbers are learned by schoolchildren. In medieval Europe, when Roman numerals were used, the division of numbers was a subject for college courses.

A mystery remains if you believe that human beings learn efficient representations on their own. But this moves the debate from "thinking" into the areas of representations and learning. We do not know what constitutes the a priori representation in the human mind that allows us to learn these kind of things. One of the great unsolved difficulties of Chomskyan linguistics is to figure out what explicitly constitutes the inborn knowledge of language. On the one hand, this knowledge is specific enough so that every child is capable of learning a human language, while no animal can do this. On the other hand, this knowledge is generic enough so that a Japanese child born in America learns English, or a Madagascarian child brought up in France learns French.

The debates about the possibility of thinking computers began with the beginning of AI and likely will continue for a while in various forms. Before turning to the contemporary state of these debates, let me briefly summarize the main points of discussions throughout recent history. The stage for the debate was set by a famous Turing test, according to which a computer can be deemed capable of thinking if a human interrogator cannot determine who answers his questions, a computer or human. During the 1980s, the rule-based AI was very popular and the arguments, both pro and contra, were often formulated in terms of rule-based intelligence. It is worth remembering that Aristotelian logic, which is the basis for the rule-based intelligence, was considered for centuries to be the main ingredient of human intelligence. Both sides of the debates included mathematicians and philosophers. The proponents included Newell and Simon (1982), who stated that logical rule systems are necessary and sufficient conditions for general intelligence, including human intelligence. Among the opponents, in addition to Searle and Penrose quoted above, I will mention Dreyfus, who emphasized limitations of logical rules and a need for recognition based on similarities. Also, Horgan and Tienson (1989) argued against the logical rules, emphasizing the ever increasing complexity of rule systems and that, instead, intelligence is related to satisfying a large number of soft constraints. (The word *fuzzy* used throughout this book is similar to their *soft*.)

Related to this debate is a discussion in Franklin (1995) emphasizing varying degrees of thinking: a computer able to solve a simple equation has a nonzero degree of intelligence. And the degree of understanding is related to the complexity of a system and to the richness of connections of a given concept to the entire body of knowledge available to the system. A similar view was emphasized by Minsky (1985): intelligence is the result of the operation of

a large number of agents, each possessing no intelligence. A conclusion seems to be that there is no qualitative difference between intelligence and nonintelligence. This book presents a contrary point of view, that there are specific characteristic properties of intelligence. I formulate the main ingredients of a thinking process, which are present even at the most elementary level. These ingredients were discussed in Section 1.1.4 for the agents of the intelligent tracker: internal models, similarity measures between the models and the world, and adaptation mechanisms. Even the simplest adaptive organisms exhibit thinking processes with these three ingredients, and in addition, the same ingredients are essential for the analysis of the human thinking.

To many people, analyses of the simplest animals and even viruses in the same terms used for human thinking may seem superficial and preposterous. Therefore, let me make a brief comment on the nature of the scientific method. When analyzing simple animals, we do not have to use complicated notions such as internal models. We can directly analyze their neural structures and describe them in terms that have little resemblance to human intelligence. It is, however, advantageous to see what is similar, so that we can understand the nature of our intelligence, as it evolved from the simplest forms. A similar situation exists in every branch of science, for example, one does not have to see the same notions of mass and force explaining the motion of planets and everyday earthly objects. Seeing similarities throughout tremendous scales of events is the essence of science. Another aspect of the scientific method is that these similarities should not be superficial and merely metaphorical, but should lead to formalizable, preferably mathematical descriptions.

Let us summarize the above discussion of thinking and understanding. Understanding refers to existing concepts (models) in our mind and to interconnections and interrelationships among these concepts. Thinking refers to arriving at understanding, which includes recognizing instances of concepts and establishing connections among them.

When our mind establishes that a subset of data in the world corresponds to a concept in the mind, it is called perception. Let us reiterate, any structure in the world that we perceive does not exist in the world alone, it is imposed by our mind. In other words, perception can be described as a process in which a bunch of photons (or a sequence of changing air pressure) suddenly "makes sense," forms a recognized pattern: a moving dot, or line, or specific sound.

A similar process occurs in cognition: a bunch of disparate concepts "suddenly" comes together and "makes sense," in other words, makes up a structured body, a new concept. Cognition refers to recognizing that one concept is a particular case of another more general concept. Or more generally, cognition refers to establishing a "super-concept": recognizing a relationship among several previously known but disparate, unrelated concepts. The nature of perception and cognition is similar. In both cases, a fuzzy uncertain a priori model that was dormant in our mind suddenly is activated by input signals, the model imposes its structure on these signals, and in this process it becomes less fuzzy and more structured, as if a resonance occurs between the input and the model. Suddenly, the model "snaps" on its input, a new structured representation emerges, and a new concept is formed. In the case of perception, this process refers to input signals coming from the world and to relatively less-adaptive, low-fuzzy, a priori models. Examples of internal models employed in perception include neural structures in the retina and in the early vision cortex (ganglion cells in retina define receptive fields for edges or moving dots, etc.). In the case of cognition, input signals are coming from other parts of the brain and a priori models could be more fuzzy and more

adaptive. A mathematical description of this process developed in the following chapters suggests that learning and creativity are related to the existence of a hierarchy of more and more general a priori fuzzy concept-models.

A similar description of perception is given by Grossberg's Adaptive Resonance Theory neural network. And the first description of this process ascends to Aristotle, in his description of the interaction between Forms of mind and matter. As far as we have arrived at a rational explanation of what are thinking processes, we can write computer codes capable of thinking. The intelligent tracker discussed in Section 1.1.4 implements this perception–cognition process. Its mathematics is developed in Chapter 4.

2.8.3 Rule-Based AI vs. Connectivism

In the midst of contradictions discussed above, heated debates revolved around the question of how important parallel organization of computations, inspired by the neural networks of the brain, is for mathematical theories of intellect and whether artificial intelligence could be founded on sequential computations utilized by common (von Neumann's) computers. I would like to note three aspects of the controversy between the parallel and sequential computational methods. First, mental processes in the brain could involve an as yet unknown physics of interaction between quantum and macroscopic states (Penrose, 1989), which is not equivalent to the Turing computer and *is not reducible to an algorithmic* mathematical description. Second, rule systems of the Plato–Minsky type could be implemented using parallel neural-network-type computers within the frameworks of existent computational concepts. And third, investigations of the parallel organization of the brain computations may lead to a new physical intuition of mind and creation of new mathematical methods.

Most of the neural computational concepts introduced so far are equivalent to the Turing computer and can be modeled on a common, sequential computer. Still, quantum computation may have an important role in the brain, and may surpass the fundamental limitations of the Turing computers. If the role of quantum computation in the brain will be confirmed by further research, this will open a most intriguing aspect of the problem considered by Penrose, today still only a hypothetical one. We will continue investigation of this intriguing proposal in Chapter 8, where we introduce a quantum implementation of the modeling field theory, and in Chapters 11 and 12, where we consider its relationship to the Gödel theory and its more far-fetched consequences for the consciousness.

As for the second aspect of the problem, parallel realization of systems of rules within the frameworks of existing computational concepts is obviously possible, as discussed in detail in Rumelhart and McClelland (1986). Nevertheless, parallelism by itself does not solve the mathematical problem of combining apriority and adaptivity, because parallelism results in only linear computational speed up as a function of the number of processors, while the number of required computations in utilized methods grows combinatorially. That is, computational needs grow much faster than capabilities of a parallel computational system. Most of the methods used today are based on the concepts considered above; variations of the MHT algorithm remain the only widely used method to combine apriority and adaptivity, and I'll repeat again the conclusion of a recent review that "much of our current models and methodologies do not seem to scale out of limited 'toy' domains" (Negahdaripour and Jain, 1991).

With regard to the third aspect of the controversy, it is quite clear today that the neural organization of the brain does provide an inspiration toward developing physical intuition

about the mind. This should not be misconstrued as an automatic endorsement of all "neural" paradigms. But it is clear that our own mind is still the best intelligent system we know, and studying principles of its material organization in our brain is certainly helpful and inspiring as much as studying motion of planets was for the development of physics of the material substance.

2.8.4 Emerging Debates

Mind as Thinking vs. Mind as Acting. One aspect of this topic is related to the old debate of declarative vs. procedural knowledge (know *what* vs. *how*). However, it moved into the new context: actions themselves rather than knowledge about actions. Traces of old disagreements still can be perceived. Production systems choose actions among a predetermined set of alternatives. And the adaptivity of actions is achieved in production systems due to planning, which evaluates multiple combinations of elementary actions (see discussion of Soar in Section 2.4.2). Contrary to this, Grossberg and his co-workers study neural mechanisms of actions. These include specialized neural networks for coordinated control of multiple muscle groups and for sensorimotor coordination. It turns out that these control mechanisms are similar to parametric models: a parameter or a small set of parameters controls the overall scale of motions, whereas motions of individual muscles and muscle groups relative to each other remain relatively constant. Adaptation in these systems acquires a "holistic" character: a change in one parameter automatically leads to a coordinated adaptation in multiple muscle groups. Thus, it seems that there are behavioral concepts similar to object-concepts in the following way. Object recognition or formation of object-concepts is the result of adaptation to the world of the a priori models (object-models). Similarly, the choice of appropriate behavior is the result of the adaptation to the world of the a priori behavior-models (leading to behavioral concepts).

There are several other aspects of this debate. One proposal is to consider acting similar to thinking: some external stimuli produce concepts, other produce actions. Or is it even possible that actions are concepts? And complicated internal representations of concepts separate from actions are not really needed for most purposes. Let us remember that the brain evolved as a control mechanism for bodily actions, therefore actions are the primary manifestation of intelligence. Advocates of this point of view usually consider concepts as conscious logical statements and associate them with the frontal cortex lobe activity. They point out that a lot of actions are produced without involvement of cortex or consciousness. My view is that there is an advantage to discovering unifying organizational principles relating thinking-concepts and actions, whether the cortex or consciousness is involved or not. Concepts, or internal models of the world, are not necessarily conscious or limited to cortex, but are used in every system involved with perception (for example, a visual perception of an edge, or moving dot). Whereas behavior also involves internal models, these are adaptive internal models (representations or concepts) of actions, not of the world. And I think that every behavioral act involves both internal representations of the world and internal representations of actions (even if these are as simple as a change of a single chemical or "prewired" muscle connections).

Still another aspect of the role of behavior in the thinking process is that adaptive thinking is impossible without at least one type of action: adaptation of the internal models to the changing world. In Chapter 10 I discuss that this type of action is responsible for accumulation of knowledge and for other higher mental abilities.

Signs vs. Symbols. The sign-nature and symbol-nature of mind is the ground for the next emerging debate. The rule-based AI was historically called "symbolic AI," I believe, for two reasons. First, because abstract mathematical notations used in algebra, formal logic, and predicate computations are conventionally called symbols. And second, because other, more complicated and even mysterious meanings of the word *symbol*[9] were not appreciated and not considered at the beginning of AI. Since the end of the nineteenth century, signs and symbols have been studied by semiotics. An emergent field of mathematical semiotics is developing mathematical methods suitable for the properties of signs and symbols and for the processes of their formation and interpretation in intelligent systems. This development is closely related to the nature of the internal representations, and it is bound to have a profound effect on the entire field of artificial or computational intelligence.

Signs, Representation, Mind, and Molecules. The debates about the role and nature of internal representations still continue, taking new turns and opening new vistas. It is a very active topic in an Internet discussion of the Architectures of Intelligent Control Systems (where I regularly participate). In this interdisciplinary discussion, engineers, mathematicians, and physicists often argue with semioticians and philosophers. The discussion of representations naturally spills into the discussion of the nature of intelligence. A discussion of internal representation vs. situated behavior turns into discussing if a tornado is intelligent? A tornado is a complex nonlinear process that is adaptive to its environment. So why not consider it intelligent? Opponents of this view ask: what is the purpose of considering a tornado as an intelligent being? One opinion is that it is useless and detracts from understanding of intelligence: first, hydrodynamics is sufficient for describing tornadoes, and second, nothing useful for understanding of human intelligence will come out of it.

Another opinion is that analyzing simple systems will help understanding the gradual evolution of intelligence and its mechanisms. My view is that if we can analyze a tornado in such a way that it enlightens us about human (or at least animal) intelligence, such an analysis is useful. An extreme view in this direction is that even if we can analyze a single simple molecule using the terminology usually reserved for intelligent beings, we will make a step toward understanding intelligence and its evolutionary emergence. This view is maintained by Taborsky (1998, 1999), who applies semiotical concepts to the world of matter. In classical semiotics, a sign is defined as something that is (or can be) interpreted by another system to mean something else. For example, a text is a sign: it is interpreted by the reader to mean something very different from ink marks on a paper. Another "less" intelligent example: a red round object on a tree is a sign: it is interpreted by a monkey as sweet juicy food. Taborsky pushes this much further: a molecule is a sign: it is interpreted by another molecule as something with which it can interact. Note the relationship of this discussion to representations: in all three cases interpretation is accomplished through internal representations. In the case of a molecule, the internal representation is a particular active site on a molecule that interacts with another molecule.

Taborsky maintains that matter exists only in an organized form. There is no unorganized matter. Organized matter exists only in interaction. Therefore everything is a sign and everything is an internal representation.

What do you think? Does it sound a little far-fetched? Could it possibly be related to intelligence? Meystel thinks that it is a waste of time, and that nothing useful could come out of this discussion for the design of intelligent robots. My first opinion was even worse. Was

the entire scientific method not made possible only after Descartes separated matter and spirit? Before that, the spiritual emanation of Neoplatonics penetrated the entire material world and interfered with the very notion of causal laws. But Taborsky made me change my view. We made a long journey from medieval philosophers toward understanding of matter. Spiritualization of matter that fascinated Neoplatonics and alchemists is not a threat to our consciousness any longer (or so she thinks). Our aim is "to materialize" the spirit, to come up with a scientific explanation of the mind. And using terminology, previously reserved for intelligent beings, to describe matter is a step in this direction. I am ready to declare myself a convert. I feel that the discussion of a molecule as a sign will affect my future designs of the architectures of intelligent systems.

Hierarchy vs. Heterarchy. Hierarchical organization is needed for many different reasons and one of them is fighting the combinatorial complexity. Research in hierarchical architectures pursued by Albus and Meystel will be discussed in Section 2.11. Grossberg cautions them along their way. He presents a number of arguments based on cognitive–psychological experiments that strict hierarchies are rarely maintained in the brain. First, the brain is made up of many modules often acting in parallel. Second, even within a single module, the overall architecture of which is known to be very much of a hierarchical nature (such as visual cortex, or speech perception), feedback loops are clearly present. (For example: in vision, perception of local cues is affected by the global picture; in speech understanding, perception of phonemes is affected by the meanings of words.) It seems the situation is clear in this regard: there is less need for discussions of opposing views as for consciously accounting in engineering designs for the feedback loops as needed. The architecture of intelligent systems is a heterohierarchy. Another aspect of this topic, which is less clear, is designing architectures with fixed, predetermined hierarchical levels and fixed predetermined heterarchical loops vs. adaptive, self-organizing systems, in which multilevel heterohierarchy emerges in the process of learning. Possible mechanisms for such self-organization are not yet clear.

Emotions. The nature of emotions remained a mystery to the rule-based AI. Psychological difficulties compounded the difficulty of the rational analysis. On one hand, the nature and potential complexity of a "machine," as partially revealed by Gödel, remain misunderstood by both philosophers and scientists. On the other hand, the works of Kant, Freud, and Jung initiated a rational analysis of the psychological nature of emotions and consciousness, but this aspect of their work remains misunderstood. Until recently, the subject remained a mystery, covered by introspection unaided by rational analysis. In the 1990s, researchers of neural networks began an exploration of emotions and consciousness in neural circuits. The field was pioneered by Grossberg, who was far ahead in this field, when in the 1970s he began to study emotional or affective signals as an integral part of learning and behavior. In Grossberg's networks, affective signals have an evaluative role: they influence formation of internal representations and recognition of objects and situations according to their importance for satisfying instinctual needs.

A different approach to understanding emotions is advocated by Gudwin, which is based on the classical semiotical notion of three types of signs: designative signs, appraisive signs, and prescriptive signs. Designative signs are concepts. Prescriptive signs are actions. And appraisive signs are emotions. Gudwin sees no difference between the three types of signs, the only difference is how they are used inside the intelligent system. I have

a problem with this approach. The problem with treating actions similar to thinking was discussed above. The problem of treating emotions similar to the thinking-concepts is that it potentially leads to combinatorial complexity. This was the case with Soar (Section 2.4.2): Soar treats evaluative signals (preferences) as logical rules, similar to thinking-concepts. A combinatorially large number of productions should be generated and evaluated.

Emotional, evaluative signals are related to Kant's ability for judgment: it is an ability for evaluation with respect to understanding. Judgment influences formation of internal representations and recognition of objects and situations according to their importance for satisfying the instinctual need for understanding. It follows that we possess the instinct for understanding and knowledge. I am not aware if such an instinct was previously identified by psychologists or biologists, and I do not know if my conclusion will spur a new debate. I will further discuss this instinct and its mathematical forms in Chapter 10.

Consciousness. Scientific discussions of consciousness are just beginning. Everything is debatable: is it "real" or just an epiphenomenon, something that is felt, but does not really affect the working of the brain? Why is it needed? How did it appear in evolution? When and at what level of organism complexity does it appear? Are dogs conscious? Are worms conscious? Where is it located in the brain? What is the mathematical difference between conscious and unconscious? What is the relationship between collective and individual consciousness? Do we really possess individual consciousness? Chapter 12 is devoted to this debate and to the answers suggested by modeling field theory.

Emergent Debate Conclusion: Unified Science Is Emerging. The future progress toward understanding intelligence and building intelligent systems will require and will greatly benefit from unifying the understanding and knowledge about intelligence and mind accumulated in several fields, including "hard" and "soft" sciences: engineering, mathematics, physics, psychology, neurobiology, semiotics, and philosophy. This will require the difficult job of coming up with unified terminology and reconciling contradictions. Many of these contradictions are due to entrenched, long-existing viewpoints. Fighting against entrenched positions is never easy.

No less is required than to bridge the schism between mathematics and physics on the one side and philosophy and semiotics on the other, while including in this unification process several branches of science. Is it possible, given the fact that even within the mathematics of intelligence there are so many subfields, with researchers barely understanding each other. There are hopeful signs: Grossberg's department at BU teaches a curriculum unifying mathematics with psychology and neurobiology; several volumes appeared unifying mathematical methods of intelligence: neural paradigms, pattern recognition, statistical learning theory, etc. Differences between hard sciences and philosophy were so great that even a need for any reconciliation seemed to be out of the question. The hope is in that this process has started. The benefits are impossible to overestimate.

2.9 Society of Mind

2.9.1 Society of Agents

In *Society of Mind* (1985), Minsky powerfully articulated a concept of mind composed of multiple semiautonomous agents, each performing its own task, each having very little

intelligence, if any. This concept exerted significant influence on the entire AI community and is now widely accepted. Minsky defined intelligence as "mental skills that . . . we admire but do not understand." This definition emphasizes his belief that there is nothing specific about intelligence, no specific element of mind, nor architectural organization, nor mathematical technique that specifically is responsible for intelligence.[10] The idea that such specific elements of intelligence can be identified he called a myth. (In the present book we will find the "elements of intelligence" that have eluded Minsky's search; we identify the specific elements of intelligence embodied in every intelligent agent, in every "atom" of intelligence, the general principles that are uniformly operational in simple organisms and in simple acts of cognition as well as in cognition of complex concepts.)

Minsky's book was influential in moving many scientists in the area of rule-based AI toward combining rule-based and connectivist architectures, toward a need to consider mind operations at "subsymbolic" levels, and toward considering architectural hybrids. He wrote, "Logical thinking. . . . The popular but unsound theory that much of human reasoning proceeds in accord with clear-cut rules. . . . In my view we employ logical reasoning only in special forms of adult thought, which are used mainly to summarize what has already been discovered. . . . Most of our ordinary mental work . . . applying . . . our representations of seemingly similar previous experiences." Minsky, however, did not develop measures of similarity different from the logical ones, nor emphasize the need for such measures. Chapter 4 identifies the measures of similarity as a fundamental aspect of intelligence, and develops their mathematical techniques.

2.9.2 Types of Agents

Minsky discusses a number of specific types of agents. *Sensors* are sensitive to signals coming from the world outside the brain. *Demons* constantly watch for certain conditions and act when they occur. *Recognizers* identify specific signals. *Nomes* is a whole class of nonadaptive agents including pronomes, isonomes, and paranomes. *Paranomes* are agents that affect in parallel and in a similar way agencies operating in several different mental areas. They operate by activating signal-agents called *isonomes* that have a similar effect on several agencies. *Pronomes* are agents associated with a particular aspect of representation, for example, they attach other agents to each other. *Polynemes* are adaptive agents that learn from experience to arouse different activities in different agencies. Minsky discusses, in particular, a number of agents related to memory. *Script-agents* generate a quick automatic sequence of actions; they gain in speed but lose in adaptivity: their sequence-actions are relatively nonadaptive. *Nemes* are agents that generate a fragment of an idea; every agent operates within a context established by a collection of nemes connected to it.

One of the old questions about memory is how storing and using memories interact. Minsky developed a memory theory called K-lines that addresses this question. K-lines are memory-agents that reactivate particular groups of other agents. Say, a memory-agent "St. Petersburg" reactivates the whole set of agents related to this experience: palaces and museums, channels and bridges, movies with tsar balls and bolshevik revolution, and the states of your mind during these movies, or during your visits there. K-lines partially return your mind "back there" by activating agents that were active at the time. Thus, K-lines store memory in their connections to many other agents, and this memory is used by activating all these agents at once.

2.9.3 Frames and Unity of Apperception

In the mind as a society of agents, the communication between agents becomes the crucial point of the society organization. How do agents agree among themselves about a unified view of the world, and how is this unified view maintained? And even more specifically, how does a large number of agents form a unified picture of the surrounding so quickly? "When we enter a room, we seem to see the entire scene at a glance." This property of mind that Kant called the "unity of apperception" is achieved in Minsky's theory by special kind of agents that he calls *frames*. A frame is a structure that is, a priori and in part, acquired from previous experience. Frames have "blank fields" or terminals; for example, a house-frame has terminals for rooms; a room-frame has terminals for windows, doors, furniture, etc. The mind has millions of frames representing stereotypical situations. Terminals are the connections at which other types of information can be attached. The terminals are filled by defaults from previous experiences. Frames maintain the unity of apperception, and defaults create an impression that we "see the entire scene at a glance."

As we "look around," defaults are displaced by information about the actual objects around us. The agents recognizing actual objects and memory-agents attached to them get attached to frames. Any kind of agent can be attached to a frame terminal: a K-line, polyneme, isonome, script, or another frame. In particular, pronomes attach terminals to other frames. Agents in Minsky's theory are closely related to what we call models in other parts of the book, and throughout the book we will elaborate on this relationship. In particular, frames are models—internal structures representing the world.

2.9.4 Limitations and What Is Next

Minsky is one of few scientists possessing the bravery and insight needed to delineate the limits of his theories. I mentioned in Section 2.4 Minsky's early prophetic assertion that rule-systems do not explain learning and adaptation.[11] Similarly, in *Society of Mind* he discussed limitations of rule-based approaches to building agents. Nevertheless, multiple attempts to build agents using rule-based concepts are being undertaken; again and again they face difficulties similar to those discussed in Sections 2.4 through 2.6. Although many researchers still believe that logical rules are sufficient to explain intelligence in its most important aspects, other become pessimistic about the possibility of understanding intelligence at all. With regard to this undue simplification on the one hand and pessimism on the other, I would like to list concepts that Minsky calls myths:

Consciousness as self-awareness about mind processes*
Creativity as a distinctive form of thought*
Intelligence as a specific element of mind*
Introspection as an ability to apprehend directly the working of mind†
Intuition as an ability for an immediate perception of truth†
Metaphor as distinct from thought
Self as a special part embodying the essence of the mind*
Freedom of will as distinct from either causality or chance†

In this list, a dagger (†) indicates a topic that is further discussed in the current book and an asterisk (*) indicates a topic on which a substantially new view is developed in the current book.

Let me mention here two of the topics whose treatment in the *Society of Mind* is fundamentally inadequate. Minsky considered emotions to be separate agents like many other agents, including those responsible for recognition, memories, and behavior. Minsky called special agents involved in emotions protospecialists; they are responsible for instinctive behavior. In other word, there are no difference in principle between emotions and concepts. In this way, emotions are like preferences in Soar. We discussed that this is inadequate mathematically. Treating emotions like concepts is one of the main causes for combinatorial explosion, it contradicts psychological understanding of roles of emotions and concepts in our psyche, and it contradicts the philosophical analysis of the roles of emotions and concept due to Kant. This misunderstanding of the roles of emotions is related to another fundamental limitation of Minsky's theory, that there is nothing specific about intelligence, no specific element of mind or specific mathematical technique required to describe it.

The present book makes a next step in understanding these complex issues. The opposite idea is developed in the following chapters, that there are specific elements of intelligence. They involve a dynamic process of concept formation in which fuzzy a priori concepts interact with input signals to form new concepts, which are crisp or less fuzzy than the a priori ones. This process of concept formation employs a mechanism of interaction between concepts and emotions. Mathematical description of this process is developed throughout the book (mostly in Chapter 4). This mechanism is uniformly employed in simple organisms and in the human mind, and at multiple levels of a heterohierarchical organization: in perception (formation of percepts from sensory signals) and in cognition (formation of new concepts from previously learned concepts). The heterohierarchical organization as well as a variety of a priori concepts and experiences determine the richness of the intelligence. An appreciation of these mechanisms as fundamental elements of intelligence is helped by relating the mathematical concepts to the concepts in philosophy, semiotics, and psychology. This relationship is interspersed throughout the book, mostly in Chapters 3 and 10.

2.10 Sensor Fusion and JDL Model

2.10.1 Sensor Fusion and Origins of JDL Model

Research and development laboratories of the Department of Defense, through their arm called the Joint Directors of Laboratories (JDL), since 1986 sponsored the Data Fusion Group that was directed to codify and standardize the technology of data fusion. For historical reasons, fusion technology addresses a number of functions associated with intelligent systems. This includes combining diverse sources of information into a unified picture of the world, managing these resources, and reasoning about the importance of various pieces of information. Fusion technology has evolved in various military applications, such as automatic target recognition, noncooperative identification friend or foe, surveillance. The Data Fusion Group produced a model of the sensor fusion process, called the JDL sensor fusion model, which served as a basis for the standardization effort. The JDL model summarized results of fusion efforts evolving through a number of years and inevitably it has a tilt toward approaches popular during these years; in particular, its overall architecture is centered around a symbolic-AI paradigm. Nevertheless, it is intended as a flexible model, and it continues to evolve, providing room for alternative algorithmic and neural network approaches.

2.10.2 Definitions, Issues, and Types of Fusion Problems

Fusion of information from multiple sensors and other sources provides more information than a single source or sensor. Fusion is used to solve problems that cannot be solved by using individual sensors. Humans and animals routinely fuse data from multiple sources. Among applications of fusion technology are military as well as commercial problems: Internet and database searches, data mining, remote sensing for agricultural and environmental data, cartography, mineral and petroleum exploration, well logging, military surveillance, automatic target recognition, mine detection, robotics, guidance of autonomous vehicles, security systems, medical diagnostics, and monitoring of industrial manufacturing processes. Sensors utilized for multiple sensor systems include passive and active sensors utilizing various bands of the electromagnetic spectrum, passive and active acoustic sensors, etc. Other information sources include data and knowledge bases, communication messages, and human operators. A tremendous need exists for information fusion in computer networks, especially Internet and accessible from Internet databases.

Various aspects of fusion are summarized in Table 2.10-1. Complexity of a fusion problem is determined most of all by how difficult it is to associate data and objects. In a simple case, data are associated with specific objects, and multiple sources are used to accumulate evidence about a specific object or event. For example, a camera records intensities of three colors (three spectral bands), red, green, and blue (R, G, B) for every pixel. So, three color intensities are associated with every pixel. Fusion of this information for pixel classification can be accomplished by using three-dimensional classifiers, or three-color models, using *classification* techniques described in Chapter 5. Similarly, records in different databases could be referenced to a specific object or event (for example, medical treatment records are referenced to the patients), so their fusion again is reduced to classification in multidimensional spaces. More complicated cases require *association*, when sensors are located in different places (two cameras observing a factory floor), or when records in

TABLE 2.10-1
Various Aspects and Complexities of Fusion

	Simple	Complicated	Very Complicated
Example	Combine data or measurements for a given object (pixel, or event)	Combine data or measurements when multiple objects (or events) are present	Combine data or measurements when multiple objects (or events) are present
Type of information used as a basis for fusion	No specific information is needed for fusion	Geometrical (3-D location and motion)	World or situational understanding
Technique	Multidimensional classification (Ch. 5)	Association and classification (Ch. 7)	The entire field of computational intelligence
Additional functions		Direct sensors or database searches	Direct sensors or database searches (attention, behavior generation)
Type of internal model	Statistical	Statistical + (dynamic, geometric, etc.)	Hierarchical, multiresolutional "world model"

various databases have no clear-cut association rule nor index as to which records should be combined. Complex cases require the development of a unified picture of the world (or situation) in order to properly associate information from various sources, such as for making complex business or investment decisions, or even to move successfully around the room, performing simple everyday tasks. Fusion addresses not only combining readily available information, but also directing sensors or database searches in an efficient manner, based on currently available information. Broadly understood fusion encompasses a significant part of intelligent system functions.

The column "simple" in Table 2.10-1 does include very complicated problems, such as high-dimensional classification or image recognition discussed in Chapter 5. But their complexity is not related to fusion specifically. The column "complicated" involves problems in which association of information from various sources is nontrivial; still they could be solved by using relatively simple object models of the types considered in Chapters 5 through 7. The last column in this table addresses a major part of the entire field of "intelligence," which we discuss in Chapter 10.

Fusion is greatly simplified if a few objects or events of interest can be first detected using individual sensors/records, and fusion is required only as a last step for improved identification of objects. The problem becomes more complicated if fusion has to be performed at an early stage in processing; for example, when not only classification, but also detection and tracking of objects of interest cannot be accomplished with individual sensors, or when information of interest can be identified only when several pieces of data from different databases are brought together. Fusion always improves performance in simple cases due to additional information available from multiple sources. But in complicated cases this is not automatically so; this is discussed further in Section 7.4.3.

2.10.3 Sensor Fusion Levels

Fusion is most simple when preliminary decisions can be made using individual sensors, and decisions from several sensors are fused for confirmation in complicated cases. This is called decision-level fusion. When individual sensors cannot provide for reliable decisions, more complex procedures are used. Fusion can be performed at the level of data, or, alternatively, some degree of information extraction can be first performed for each sensor. Extracting information from individual sensor data simplifies the fusion, but some information may be lost in the process. Combining all the data from multiple sources before making a decision potentially provides more information, but is computationally complex, often leading to combinatorial explosion. Therefore, data-level fusion is often accomplished by extracting features from multiple sensor data and designing classifiers in the feature space.

To overcome the high dimensionality of data-level fusion, feature-level fusion is used. In this approach, features are extracted from individual sensor data and fusion of multi-sensor data is accomplished by combining features. This is the most common approach to fusion. The success of this approach depends on feature selection that should be based on understanding of the physics of the observation process and phenomenology of object identification as discussed in Section 1.3. In model-based fusion, data-level fusion requires models of data, which model pixel (or sample) values for individual sensors; feature-level fusion requires models of features, and decision-level fusion requires models of decisions for individual sensors.

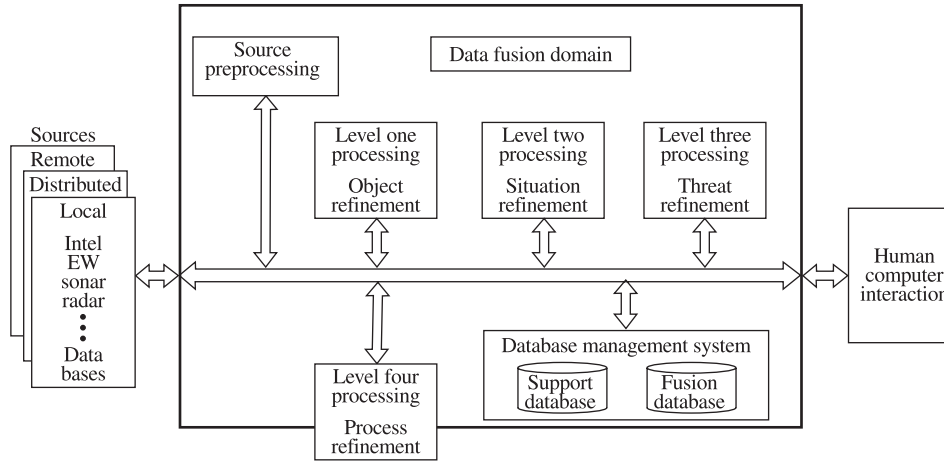


Figure 2.10-1 JDL data fusion model.

2.10.4 Hierarchy of JDL Model Organization

The JDL model identifies the most important functions and processes of a multisensor fusion system and organizes them in a hierarchical system. These include access to data sources, source preprocessing module, four processing levels, a database management system, and a user interface as shown in Fig. 2.10-1. This division into a hierarchy of processing modules or levels is somewhat artificial and is not always easy to accomplish in practice: in real data-fusion systems, partitioning data among a strict hierarchy often cannot be achieved and these processes are integrated. A source preprocessing module performs data prescreening and initial allocation of data to appropriate processes (for example, raw sensory data are allocated to level one processing, whereas attention-requests or alerts are allocated to level three processing). Level one processing combines and refines information on individual objects. At level two, relationships among objects are established in an environmental context. Level three derives a unified picture of the world, projects the current situation into the future, and draws inferences utilizing knowledge bases. Level four monitors other processes and allocates resources. Techniques and examples considered in Chapters 5 through 8 are mostly relevant to the first, second, and fourth levels. Level three deals with interpretation of information on a higher cognitive level and requires more complex models of environment, situation, etc. A hierarchical organization that is present in an incipient form in the JDL model is an important general organizational principle for intelligent systems. It is considered in the next section.

2.11 Hierarchical Organization

A concept of hierarchical organization was proposed to overcome the combinatorial computational complexity of rule systems and model-based systems. In a hierarchical organization,

the entire problem is decomposed, first, into multiple resolution levels, and second, each level is decomposed into limited-scope modules. The rationale is that every module has to deal with only a limited set of distinguishable objects and alternatives, so that combinatorial searches even if required do not lead to the prohibitive combinatorial explosion. Also, combining multiple levels can be used to achieve the required detail of the analysis, plans, and actions.

A hierarchical architectural model of intelligent systems has been under development by the National Institute of Standards and Technologies (NIST) since the late 1970s. Currently, its main developers are Albus and Meystel. This section overviews the NIST architectural model and summarizes the views on intelligence of the system developers. The NIST model aims at the general theory of intelligence encompassing biological, machine, and social intelligence. It incorporates the knowledge gained from many sources, including artificial and natural systems, brain organization, psychology, psychophysics, neural networks, and experience accumulated in robotics research, control theory, and industrial automation development. Intelligence is defined as "the ability of a system to act appropriately in an uncertain environment." The appropriate action is that which increases the probability of the achievement of the system's goal. The goal is defined external to the intelligent system. For artificial systems, goals are defined by designers. For biological creatures, it is gene propagation. Albus and Meystel differentiate between adaptation and learning, using the adaptation term for intelligent functioning in a changing environment and learning for becoming more intelligent from experience.[12] Learning is consolidating short-term memory into long-term memory, with subsequently modified behavior. Learning is a mechanism of storing knowledge about the world, acquiring skills, and knowledge of how to act. They emphasize that many creatures act intelligently by using instincts, without having learned anything.

Natural intelligence is a result of natural selection. The brain is foremost a control system with the primary function to produce goal-seeking behavior. Abstract thought is a relatively recent phenomenon in evolutionary terms. Intelligence is a mechanism of advantageous behavior generation. Higher levels of intelligence include the ability to represent knowledge in an elaborate world model, to plan before acting and reason about possible results, and to act successfully in a complex environment.

The NIST intelligent system is composed of four main elements or subsystems: sensory processing or perception, world modeling, value judgment, and behavior generation. It also has two "auxiliary" subsystems for interacting with the outer world: sensors and actuators. All these subsystems do not function autonomously, but in interaction with each other. The internal world model supports the functioning of other subsystems and the coherency of their interaction. The world model is the intelligent system's estimate of the state of the world. It includes the database of knowledge and simulation capabilities, which produce expectations for the perception system and predictions for behavior generation planning. The world model plays a role similar to the Plato's Ideas and Aristotelian Forms: it imposes the forms of perception, cognition, and action onto the outer world.

Perception–sensory processing compares sensory observations with expectations generated by an internal world model and integrates similarities and differences between observations and expectations over time and space. On the basis of the world model, perception recognizes objects and relationships and fuses sensory data into a unified and consistent perception of the world state. Thus, the world model is responsible for what Kant called the Unity of Apperception (unity of consciousness).

The world model also supports the value judgment system, which determines good and bad, rewards and punishments, important and trivial, certain and improbable. The value judgment system models emotional biological systems. It computes costs, risks, and benefits of observed situations and planned actions; it assigns believability or uncertainty to state variables and attractiveness or repulsiveness to objects and events.

Currently, the value judgment module evaluates objects and concepts that have been identified by the perception module. Relationships between perception–cognitive and evaluative–emotional functions present one of the fascinating and least understood aspect of intelligence. The separation of cognitive and emotional functions in the NIST architecture is similar to the separation existing in Soar (between productions and preferences) and is one of the primary causes of combinatorial explosion (see relevant discussions in Sections 2.5 and 2.7). Cognitive and emotional processes are more closely integrated in our brain, where emotions are inseparable from cognitive processes. We, humans and scientists in particular, are under a strong impression that we can keep our emotions separate from our thoughts, especially when thinking about scientific problems. But it is well known from psychology that this impression is false, and the psychological reasons for this misperception are discussed in Chapter 12. Our ability to perceive the beauty of a scientific theory is due to this close integration of emotions and thoughts. The roots for this conclusion ascends to the work of Kant, and its relationships to the architectures of intelligent systems are discussed in Chapter 10. Close integration with emotions is characteristic of both lower and higher level cognitive processes. Neural circuits integrating "low-level" cognitive and emotional processes were investigated by Grossberg, as discussed in Section 2.14.

The behavior generation system selects subgoals, generates plans, decomposes them into tasks, and monitors their execution. Subgoals and plans are generated by iterative interactions among the system elements: the behavior generation subsystem hypothesizes plans, the world model predicts their results, and the value judgment evaluates them. The best plans are selected for execution. The interaction of the main elements of intelligence is illustrated in Fig. 2.11-1.

The main elements of intelligence are organized in a hierarchical architecture illustrated in Fig. 2.11-2. Each node in the organizational hierarchy contains the four element-modules. There are both vertical, hierarchical, horizontal, and lateral communications among nodes, which are orders of magnitude less intensive than communications among the modules within a node. The specific configuration is not necessarily fixed: nodes could be dynamically reconfigured among vertical substructures as needed. The main feature of a hierarchical organization is that every element has multiple hierarchical levels of spatial and temporal aggregation, or levels of scales and resolutions. Behavior generation hierarchy defines temporal and spatial decomposition of goals and tasks: temporal historical traces and planning horizons, and spatial ranges of maps and controls. Similarly, world modeling and signal processing are aggregated into temporal and spatial aggregation levels, temporal resolution of events and spatial resolution of objects and maps. It is assumed that at each higher level, the temporal and spatial resolutions decrease by an order of magnitude; correspondingly, goals and planning horizons expand in scope by an order of magnitude.

A need for combining the hierarchical structure with resolution levels follows from several considerations. Hierarchically nested control loops may become unstable unless their bandwidths differ by an order of magnitude. A hierarchical structure maintains integrated short- and long-term functioning. And let us reiterate that its main aim is to eliminate

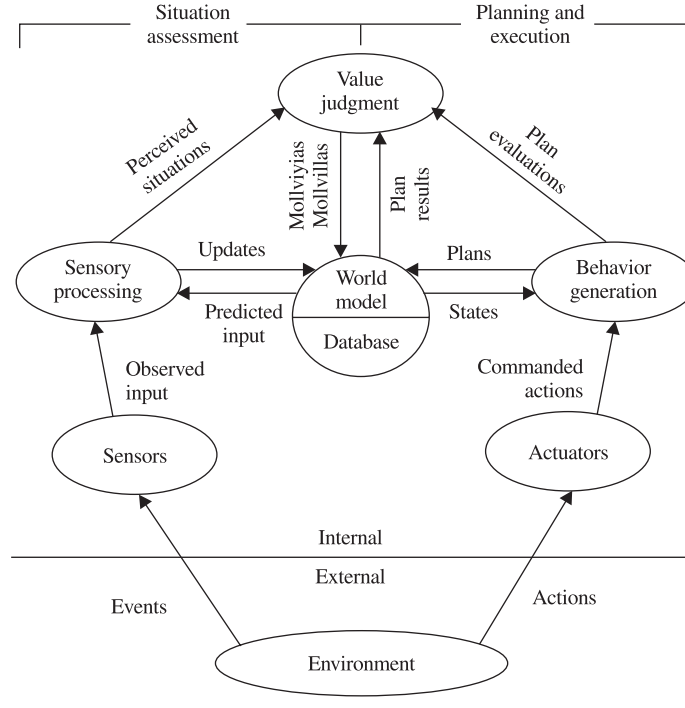


Figure 2.11-1 The elements of intelligence and the functional relationships between them.

combinatorial explosion. Reduced resolution at higher levels results in aggregation of data and models and is equivalent to employing fuzzy logic with the hierarchical structure of the degree of fuzziness. Relationships between fuzzy logic and hierarchical organization are an ongoing area of research. Zadeh proposes a combination of fuzzy logic with granularity. The concept of granularity suggests that the fuzzy uncertain concepts can be considered as "indivisible" units at the appropriate level. This implies multiple levels and hierarchical organization. According to Meystel, fuzzy logic is operational on a subgranular level, but at a higher level where the granule is considered as a unit, the Aristotelian law of contradiction (the usual logic) can be used.

This corresponds to our analysis in Section 2.1 of the fundamental role of fuzzy logic in learning. And our theory developed in Chapter 4 can be viewed as a flexible, adaptive version of these ideas. Yet we know that the brain is not a strictly hierarchical system, and it has multiple heterarchically organized modules. Even within relatively hierarchically organized modules, there are processing loops involving multiple levels. It is clear that we need to go beyond current approaches to developing hierarchical systems, which employ predetermined nonfuzzy and nonadaptive definitions of the hierarchical structure itself. This is inconsistent with the fundamental importance of fuzzy logic and adaptivity and with existing psychological and neurological data. It is contrary to the ability of our mind to create flexible adaptive hierarchies as needed, with interactions among several adjacent layers when required. It seems that the success of the hierarchical method as a general paradigm

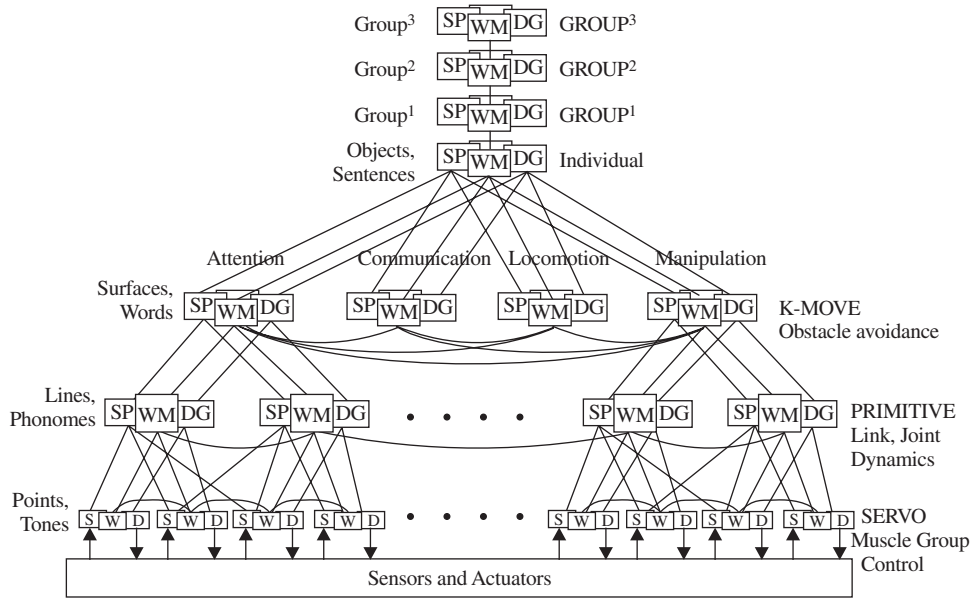


Figure 2.11-2 An organization of processing nodes such that the BG modules form a command tree. On the right are examples of the functional characteristics of the BG modules at each level. On the left are examples of the type of visual and acoustical entities recognized by the SP modules at each level. In the center of level 3 are the type of subsystems represented by processing modes at level 3.

of intelligence will require development of adaptive fuzzy methods of hierarchical segmentation. This will lead to the understanding of an emergent heterohierarchical organization.

2.12 Semiotics

Semiotics is the science of signs and symbols, an analysis of languages of all kinds, which are the meat and bread of intelligence: semiotics studies the sign-nature of mind. Its roots date to Aristotle, it was formulated as a field of study in the last century by Peirce, and was reshaped in the second quarter of this century by Morris. The founders of semiotics considered a sign as an object (called a *sign-vehicle* or *sign*), whose purpose is to refer to another object *(designatum*). This definition assumes that there is an *interpreter:* a mind or an intelligent system (not necessarily human) that interprets a sign-vehicle. In the process of this interpretation, an internal representation of an object is formed, called an *interpretant*. The process of interaction within the triadic unity of sign–designatum–interpretant is called semiosis. Morris decomposed this process into the three dyadic relationships-processes: (1) syntactics studies relations among signs, (2) semantics studies relations between signs and their designata, and (3) pragmatics studies relations between sign-vehicles and their interpreters.

Let us illustrate these concepts with a simple example. For a monkey, a red round object on a tree is a *sign* for food. The juicy sweet food is a *designatum*. Its internal representation,

interpretant, is a concept combining both these aspects; it acquires its full meaning in the interaction with other concepts and emotions in the monkey's mind (e.g., hunger, how to get it). The monkey is an *interpreter* of this sign. Syntactics includes relationships among the apple, tree, branches, other monkeys, etc. Semantics includes relationships between the visual appearance of an apple and its edible properties. Pragmatics includes relationships between the monkey and "a red round object on a tree."

Consider the embodiments of these processes. A sign as well as an interpretant ought to have internal representations. For an external sign, the representations are signals that our sensory cells receive from the world, a sensible environment. This is the first layer (the signal or "preperception") layer of an intelligent system. The structure of our mind responsible for this bottom layer is (in a gross conceptual way) similar to higher levels in the hierarchy: it has internal models, which impose structures on the input data. Interpretants formed in this first layer are internal signs, which are interpreted by the next layer, and over and over again, leading to the loop of semiosis. For an adaptive system, there are (at least) three type of loops or iterative processes involved at every layer. One loop involves a functioning of a particular sign-symbol module, which repeats the interpretation of particular signs, as they appear, and invokes appropriate actions. Inside this loop, there is an adaptation process in which a sign is transformed into an interpretant. This second loop is a part of the hierarchical ascendance loop of semiosis, which involves multiple layers of the system.

According to the basic principles of semiotics, an architecture able to perform semiosis needs to integrate three types of knowledge: designative, appraisive, and prescriptive. Every functioning loop of a particular sign-symbol module involves these three types of knowledge. Designative or conceptual knowledge is contained in the internal models, appraisive knowledge is emotional evaluative signals, and prescriptive knowledge generates behavior. The process of interpretation, the adaptation loop, involves the three ways of reasoning: deduction, induction, and abduction. It utilizes the existing knowledge to extract specific rules or models (abduction), it learns from experience (induction), and this learning is constrained by and made possible due to existing knowledge (deduction). A semiotical architecture performing semiosis is an intelligent system.

Semiotics attained the height of its popularity in the 1960s, when the computer civilization was as yet in an early stage and not ready for the complexity of semiotics. So semiotics turned to art, attempting to explain it through a formal language, which art, by itself, does not need, because art is a language. Semiotics became entrapped within the figurative metaphors of art languages; its meaning was lost, and its original appeal withered. Many scientists worked on bringing semiotics into the domain of science. Sebeok (1972) has shown that the ability to use signs and symbols is not limited to human beings; he developed Zoosemiotics, a branch of semiotics studying the role of signs in the animal kingdom. Relationships between semiotics, symbols, communication, and information were investigated by Sebeok (1977) and Eco (1976). A role of signs in natural physical systems was discussed by Wheeler (1988). Today our civilization is getting comfortable in its computer age. This process is bringing forth an important role of semiotics as a science about the sign-nature of culture, and the role of signs and symbols in the mind. An irony of the situation is that 30 years ago semiotics offered too much to the field of "symbolic AI," a field that ostensibly addressed the sign-nature and symbol-nature of knowledge, but did not possess the mathematical techniques to deal with the complexity of the subject. Both fields lost their popularity at

approximately the same time, but today, computational intelligence is ready for research into the dynamic nature of signs and symbols as revealed in semiotics.

A particular aspect of semiotics, which "symbolic" rule-based AI was not ready to address, was the complex and dynamic nature of symbols. In classical mathematics, symbols are signs or marks denoting specific, well-defined mathematical concepts, such as numbers, functions, and operations. But in psychology and in the general culture, a symbol is a word loaded with a lot of meaning. A symbol is often understood not just as a sign[13] for a specific concept, but as a complicated psychological process that involves consciousness and the unconscious, a process in which new concepts are emerging and new meanings are created. Symbols understood in this way are adaptive entities, supporting learning. In the 1960s, according to Minsky, rule-based AI was not supposed to explain learning; this task he considered at the time to be too complex for existing methods. Concepts in rule systems were static nonadaptive entities, signs designating predefined events and situations, and not dynamic symbols. Therefore, "symbolic AI" is a misnomer, a wrong name for the first attempt to represent the thought process using formal logic.

Mathematical semiotics is an emerging science that develops mathematical methods describing the processes of sign interpretation and symbol formation. These processes are closely related to the process of thought and to understanding of mind. Therefore, mathematical semiotics should bring together concepts in classical semiotics, computational intelligence, psychology, and philosophy. This task is complicated by the fact that significant differences exist among the concepts, definitions, and terminology of these disciplines. For example, the psychological notion of symbol does not correspond to the notion of symbol in classical semiotics. Instead, classical semiotics uses a word "sign" for both a sign-vehicle and the process of its interpretation, and uses "symbol" as a particular case of "sign." Throughout this book "symbol" is used according to the psychological and general cultural meaning to denote a psychological process of the interpretation of sign, and "sign" is used for a sign-vehicle.

The hope is that the mathematical apparatus will help to establish the correspondences and reconcile the differences. For example, classical semiotics considers the process of semiosis, the evolving process of the continuous formation of more and more complex signs and symbols. Meystel relates the process of semiosis evolving in an intelligent system to his concept of the hierarchical organization of intelligent systems; he concludes that semiosis is a hierarchical multilevel process, and should involve hierarchical internal representations. A mathematical theory of the dynamic symbol as a thought process is described in Chapter 10.

2.13 Evolutionary Computation, Genetic Algorithms, and CAS

What is the original source of the a priori information? Where are the a priori structures coming from? In the engineering applications of rule-based and model-based approaches, the a priori knowledge is specified by the designers. Evolutionary computation (EC) and genetic algorithms (GA) are mathematical techniques designed according to concepts of genetic evolution with the goal of explaining the evolution of internal a priori structures of mind.

EC is a broad research area developing mathematical methods and algorithms for intelligent systems and applications in many fields, which are inspired by the concept of natural evolution. Many EC systems are built of agents and the adaptation is achieved by (1) generating new agents and (2) selecting good agents and their combinations and discarding bad agents and their combinations; various appropriate measures of fitness are used for this step.

An important difference among various directions in the EC field is illustrated by GA vs. evolutionary algorithms (EA). GA operate on adaptive systems organized in two levels, genotype (a priori model) and phenotype (individual "acquired features"). Fitness is measured by the overall phenotype performance, so the performance feedback is available for the entire set of genes (genotype) for each agent, but not for each gene (allele). Also, genetic operators do not guarantee the preservation of the genotype of the best individuals. EA operate on a single-level system of phenotypes. Performance feedback is available for each feature of the phenotype, and the phenotype is considered as the "genetic information." One could say that GA model the evolution process at the individual level, whereas EA model the effect of evolution at the level of spices. A unifying theme in EC is that these type algorithms combat combinatorial complexity by accumulating past experience. This accumulation is not linear, but combinatorial: every generation "tests" not just individual genes, but, in some indirect way, all combinations of all subsets of alleles (schemata) present in the population are tested in each generation. On average, the proportion of each schema that provides for an evolutionary advantage exponentially increases in the population. In the next section we consider a GA technique developed by Holland.

2.13.1 Complex Adaptive Systems (CAS)

Holland searches for the solution of the issues of adaptivity, apriority, hierarchical organization, and combinatorial complexity in his theory of intelligent complex adaptive systems (CAS). This section briefly reviews the concept of CAS organization that includes genetic-type algorithms for learning and evolution.

CAS are composed of intelligent agents. Each agent is a "simple" if–then rule:

$$\text{if } (\mathbf{a}) \text{ then } (\mathbf{b}) \tag{2.13-1}$$

The if part of an agent tests conditions, and the then part performs actions. In a simple case, an agent is a Pavlovian stimulus–response arc: **a** is a signal detected by a sensor, and **b** is an action performed in the environment by an actuator. A powerful generalization proposed by Holland is to consider **a** and **b** as general type messages that can be received and sent by agents (**a** is an a-message receiver and **b** is a b-message transmitter). A general type message can be considered as a string of three types of digits: (0,1,#), where # means "don't care," so that, e.g., a message 011001 can be received by $\mathbf{a} = \#11\#0\#\#\#\#\#$. Holland envisions swarms of agents sending and receiving messages: some messages might come from sensors or could go to actuators, but most devote their activity to internal thinking process: building and estimating internal models.

For agents to be adaptive, there should be ways (1) to generate new rules (new a- and b-message receivers and transmitters) and (2) to select good rules and their combinations and to discard bad rules and their combinations. Holland considers genetic algorithms for

rule generation and two types of algorithms for rule selection. New rules are generated by two types of genetic operators, crossover and mutations. Crossover acts in the process of "mating" of two parental agents: with certain probability, two agents mate and produce an offspring. An offspring is a new agent with a-message receiver or b-message transmitter obtained from the parental ones by a crossover operation, an exchange of substrings between two message strings as illustrated in Fig. 2.13-1a. A crossover point along the strings is selected randomly. A crossover mechanism of the new rule generation provides for utilization of accumulated experience: new message strings are formed from existing building blocks, substrings. And, with appropriate rule-selection algorithms, useful substrings propagate to offspring more often. Mutations act by random replacing of a single character by a different one, as in Fig. 2.13-1b. Mutations are needed to retain adaptivity even within those substrings that came to dominate the population genome. In biological systems, mutations are rare events (and this has to be so, since they effect random deviations from what is known to be good).

An algorithm considered by Holland for rule *selection* is a credit assignment algorithm, which is a variant of Adam Smith's capitalistic "invisible hand." According to this algorithm, agents within a CAS system are in competition for posting their output b-messages on a "web page." They "bid" for a limited number of available slots and higher bids win. They have to pay with available "cash." Similarly, agents are in competition for using input information that they have to "buy" from the web page. Cash paid is credited to the posting agent. Every agent posting a message also posts a bid price for using his message. Price is recorded within the b-message, so it is also a subject of adaptation. An ultimate source of the cash floating around the system comes from the outer world, when a system gets "food" or other vital resource (in an engineering system, it could be any desirable performance goal, say a stock market gain). I prefer a fuzzy modification of this algorithm, according to which all agents post their messages with a fuzzy strength (weight or probability) corresponding to the usage of their messages by other agents, and utilization of messages is proportional to their relative strengths. This fuzzy modification is preferable on theoretical grounds (it adds stability and speeds up learning) and is reminiscent of neural connection weights, which are strengthened when used and weakened from disuse. This algorithm potentially leads to an evolution of the CAS system (a population of agents). The evolution is not necessarily a very stable process; there could be strong fluctuations in efficiency (similar to those that occurred in early capitalistic societies). Still, it might be expected that in competition among agents, a gradual evolution of a CAS system will occur: those agents that use "good" information as their input eventually get better rewards from the environment, they can better pay their "suppliers," and, accordingly, "good-message suppliers" get paid more and can exert a stronger influence on the system performance.

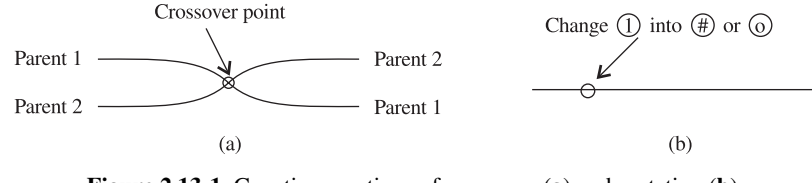


Figure 2.13-1 Genetic operations of crossover (**a**) and mutation (**b**).

Another algorithm for rule selection is a genetic selection algorithm. According to this algorithm, the probability of mating among agents is proportional to their fitness. Fitness can be determined by direct "survival" in an environment, or by the amount of "cash" accumulated by each agent according to the credit-assignment algorithm. Thus, offspring in each generation are expected to outperform the average fitness of the population. To prevent overpopulation and speed up the adaptation, lower fitness agents can be replaced randomly by the new (offspring) agents, or the life span of an agent can be limited (and related to its fitness).

In CAS systems and genetic algorithms, the unit of adaptation is not an individual agent, but a population or system of agents. Evolutionary "pressure" leads to selection of "good" building blocks or schemata. Schema is a mathematical notion corresponding to a generalized concept of a collection of building blocks (or substrings) that coevolves in the evolution process. A schema may include "ignored" positions (*); for example, schema 1#**#***** includes all strings beginning with 1# and having # in the fifth place. Schemata are not used in the algorithms, but for mathematical analysis of genetic algorithms. Mathematical analysis and computer simulations indicate that genetic algorithms and CAS system organization lead to exponential growth of the proportion of "good" schemata in the population, and therefore lead to increased fitness of the population of CAS systems.

Unsolved problems in CAS theory include insufficient understanding of how to make sure that a population of agents continues efficient evolution in complex systems with long genetic codes and with significant random deviations of payoffs from average fitness values. The difficulties are of two types: if evolutionary pressure is strong relative to other factors, the learning is fast and a system may quickly come to a suboptimal equilibrium point. This effect is exacerbated for small populations. If evolutionary pressure is weak, for a large population with randomness in individual payoffs, there might be a tendency to preserve too much diversity in the population, crossovers might break good schemata faster than they propagate in the population, and no evolution occurs.

2.13.2 CAS: Complexity vs. Fuzziness

Let us review the conundrum of combinatorial complexity vs. CAS systems. Throughout this chapter we made the point that fuzzy logic is needed to eliminate combinatorial complexity. But genetic algorithms and CAS systems, at first glance, do not use fuzzy logic: genetic code is a nonfuzzy string of characters. So, where does the resolution of combinatorial complexity come from? To answer this question, let us analyze the potential for combinatorial explosion of CAS and the means by which it is resolved. We will see that combinatorial explosion is avoided in CAS by means of fuzzy logic acting at the level of schemata.

We begin this analysis by comparing genetic algorithms to a naive adaptation approach: by trial and error. Denote the length of a genetic code by L (this is a combined number of characters in **a**- and **b**-messages). The total number of if–then rules with the L-length code is 3^L, since any of the three characters, (0,1,#), might occur at any of the L places. For moderate complexity codes with $L = 100$, the total number of rules is $3^{100} \sim 10^{48}$. This very large number, comparable to the size of the universe, is the very familiar combinatorial explosion. Although it is good that CAS systems have such a potential for diversity, it is clear that this large number of rules cannot be evaluated by trial and error. Clearly, even a very small part of this problem cannot be approached with brute-force trial and error combinatorics. The

power of genetic algorithms is that they do not search randomly, but use past experience for directing future searches and build new rules from building blocks identified in the past. The information about past experience is accumulated in the proportions of various schemata in the population. A unit of adaptation is a schema rather than an individual agent.

Compare information representation in an agent and in the population genome. An agent is characterized by its "genetic code," $(\mathbf{a}, \mathbf{b})$, which we will denote $\mathbf{g} = (\mathbf{a}, \mathbf{b}) = (g_1, \ldots, g_L)$. For the following conceptual analysis it is convenient to consider each g_i as taking one of two values, 0 or 1 (in the Holland's formulation, g_i takes three values, but since any number can be expressed in binary as well as in ternary code, these representations are conceptually equivalent). Every g-code defines a crisp logical if–then statement. Consider now a proportion of each allele (a g_i value) in the population. The proportion of $g_i = 1$ is given simply by an average value of g_i in the population. Let us denote the average by $r_i = g_i$, so that the average g-code is given by $\mathbf{r} = (r_1, \ldots, r_L)$. Now, $\mathbf{r}$ is a fuzzy and not a crisp statement about the allele values. This is even better seen from the fact that there is uncertainty or fuzziness associated with each r_i. This uncertainty can be characterized by the variance, $c_i = \overline{(g_i - r_i)^2} = r_i(1 - r_i)$ (see Problem 2.13-1). If the entire population has $g_i = 1$, the variance is zero (of course, the same is true if all $g_i = 0$). Thus, schemata can be viewed as fuzzy submodels with expected values r_i and variances $r_i(1 - r_i)$ represented by a population of nonfuzzy, nonadaptive individual agents.

I would like to emphasize the relationship between fuzziness and the noncombinatorial nature of CAS adaptivity. CAS agents are nonfuzzy and nonadaptive. CAS schemata are fuzzy and adaptive. The genetic mechanism of preferential reproduction for better fitted agents creates a gradient in the space of parameters of fuzzy schemata leading to schemata adaptation. This evolutionary gradient is in the direction of increased fitness due to a mechanism that allocates fitness (or cash) to schemata, as an average fitness (or cash) of agents belonging to each schema. The very existence of a gradient and possibility for adaptation is related to the fuzziness of the schemata.

2.14 Neural Field Theories

2.14.1 Grossberg's Method: Physics of Mind

Deliberative argumentative thinking is a small part of what our mind does. A theory of intelligence founded on the paradigm of logic, like Soar, is driven by a specific type of intuition, a mathematical intuition. Mathematical theories are governed by a type of intuition, which is abstract and related to a sense of beauty of the internal structure of a theory. Soar was initially developed as an elegant mathematical tool for solving toy problems. Only much later did a serious effort toward relating Soar to the wealth of experimental psychological data begin. Scientists are different in the type of intuition that drives the development of their theories. Physical theories are different from mathematics in that they are driven by an intuition about the structure of the world. The beauty of a physical theory is in the sense of relationship between the world and its mathematical description.

The development of the physical theory of mind was initiated by Grossberg in the late 1950s. His physical intuition approach combines mathematics with the wealth of psychological and neurophysiological data. In his early work during the 1960s and 1970s,

Grossberg developed a number of neural mechanisms that later became very popular; these include additive neural network, instar and outstar architectures, competitive and cooperative learning, and mechanisms of emotional control in recognition, learning, and attention. One significant aspect of his work he describes as a progressive "unlumping" of the lumped mathematical models, leading to ever finer identification of neural processes. But the very initial mathematical approach corresponds to the physical intuition about working of the mind.

Grossberg's method starts by identifying a minimal but fundamental realistic environmental constraint, to which a species must adapt for survival. The solution to this problem dictates a principle of behavioral organization. This principle is then formulated mathematically, using the simplest mathematical procedure adequate for the task. This general approach implements a centuries-old paradigm of the general scientific method: mathematical minimality of the solution is the Occam's razor that minimizes the number of independent principles. However, a fundamental question comes up related to the evolutionary development of our mind and brain: the prior evolutionary history may prevent the minimal solution. Thus, a mathematical solution to a particular behavioral organization principle has to be "embedded" into the properties of several principles acting together. The result is what Grossberg calls the embedding field theory. His theories are *field* theories in that they describe the collective or interactive properties of neural networks. Some of the fundamental principles of neural organization found by Grossberg and his co-workers are outlined below.

2.14.2 ART Neural Network

In the adaptive resonance theory (ART) developed by Grossberg and Carpenter, perception is a resonance between afferent and efferent signals, that is between signals coming from the outside, from sensory cells receiving external stimuli, and those coming from the inside, that is signals generated by a priori models. For example, visual perception is a process of resonant matching between stimuli and models of elementary objects of visual perception contained in the retina and visual cortex.

ART is a theoretical principle of the structure of adaptive robust feedback connections between two different levels of a neural network. One level is cognitively "higher" than the other. For example, visual cortex vs. thalamus (lateral geniculate body of the thalamus that preprocesses visual information), or visual association cortex vs. visual primary cortex. In ART, a single node at the "higher" level encodes patterns of node activities from the "lower" level. This higher level node is a concept-object: it recognizes an individual object in the lower level node activities. The lower level input pattern activates higher level nodes, the synaptic connections of which it matches. The term "resonance" refers to a coherent dynamic state that arises from matching between an input lower level pattern and a stored prototype pattern. The stored pattern is represented at the synaptic connections from the higher to lower level and is activated by a higher level node. Perception corresponds to the resonant state between the concept and the data. When a resonance occurs, both the bottom-up filter (recognized object) and the top-down prototype (concept) are updated. Therefore, the resonance is called adaptive. Also, the resonance takes a sufficiently long time (relative to the decay rates of individual cell excitation), so that the long-term memory (LTM) can be updated to store modified short-term memory (STM) patterns. The LTM update is a long-term learning mechanism in ART.[14] In addition to ART being a fundamental principle in

the perceptive and cognitive neural organization, Grossberg identified adaptive resonances in nonneural tissues. Adaptive resonance seems to be a basic design principle of the model-based development. ART is a universal mechanism for perception and cognition. It provides a mathematical framework for the Aristotelian concept of mind as meeting between the a priori Forms (internal concept-models) and matter (external stimuli-object).

The ART architecture is illustrated in Fig. 2.14-1. The input signal contains sensory data or lower level concepts (objects) that already have been recognized at a next lower level processing. It is temporarily represented at the STM in the F_1 field. This representation is affected by the input signal and by the existing expectations generated by the LTM. The LTM contains a priori[15] concept-models of a higher level. Several concept-models may partially correspond to the signal representation in F_1. This partial correspondence excites the bottom-up pathways, in which the LTM traces of these concept-models are stored. The result is an excited temporal (STM) representation of an input signal in F_2, a template of the input signal. This template, therefore, is a result of interaction of the input-signal representation and a priori internal model-concepts. It matches neither of them perfectly. But it partially matches some of the a priori concept-models and, therefore, excites top-down pathways, in

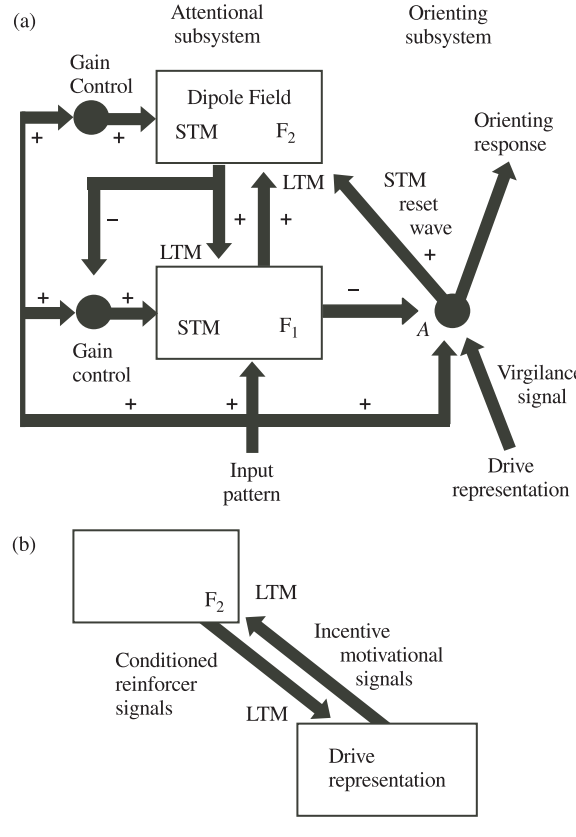


Figure 2.14-1 Adaptive Resonance Theory (ART) top-level architecture.

which the LTM traces of these concept-models are stored. Changes in excitations of top-down pathways modify the STM representation of the input signal in F_1, so that it better matches a particular a priori model-concept. A loop involving iterative modifications of F_1 and F_2 STMs may gradually lead to a high degree match between them, a match that corresponds to a particular high-level concept, and to a state of resonance between F_1 and F_2. Recognition of a higher level concept occurs when there is a match between the top-down expectations generated by a higher level concept and bottom-up input signals. If the iterative process does not lead to a match, the orienting system reinitiates the context used for the search of recognition or generates a new higher level concept. This entire process of cognition is modulated by affective (emotional) signals related to basic instincts or drives, as discussed in the next subsection.

The original ART network utilized winner-take-all competition among concept-models in the F_2 field. The winner-take-all mechanism enhances activation of the matched concept-model and reduces activation of mismatched ones, until a single activated concept remains. It guarantees stable and fast learning, but may lead to unwarranted multiplication of the number of concept-models. In distributed ART (dART) networks, matching and learning alternate between winner-take-all and distributed learning. In distributed learning, multiple concept-models are active concurrently, and neural connections are strengthened for all activated concept-models. Also in dART, traditional neural pathweights are replaced with weights that are functions of the difference between the signal and an adaptive threshold. In both these aspects, dART is similar to the modeling field theory neural networks described in this book

An experimental discovery of a resonance recognition process in cortex was made by Freeman in 1975. He found that when a cat smells an expected scent, its cortical potentials are amplified until a synchronized oscillation is achieved across the cortical region. The oscillation becomes a coherent spatiotemporal pattern. A recognized scent (an "olfactory concept") is represented as a spatial pattern of activity across cortical cells. By contrast, when a cat smells an unexpected scent, the cortical activity is suppressed.

In ART neural networks, higher level concepts compete for the evidence in the input data patterns. Competitive mathematics describes this process analogously to the Bayes equation, resulting in a probabilistic physics of choice between competing hypothesis. Grossberg hypothesized that the probabilistic physics of our psyche may explain effectiveness of probabilistic theories in many of their aspects. Probabilistic physics leads to more powerful algorithms than the combinatorial search of multiple hypothesis testing (MHT). MHT algorithms discussed in Chapter 1 were designed to utilize explicit models; they turned out to be limited by the combinatorial complexity of computations. The original ART formulation did not combine competitive learning with complex a priori models. ART's efferent patterns, generated by higher level concepts, were based on learning by remembering past experiences: by remembering previously encountered patterns or by averaging several similar patterns. The computational concept of this learning was equivalent to the nearest neighbor. This led to misunderstanding of ART as a nearest neighbor algorithm. The power of ART is in providing a general mathematical description for the interaction between two different levels of a neural network. And ART's power is fully realized when it is a part of more complex neural systems so that a higher concept-level encodes complicated patterns by combining adaptivity with complex a priori structures (models). Such more complex neural systems are being developed by Grossberg and his co-workers. Modeling

field theory developed in this book can be viewed as a further development of the theory of mind in the direction initiated by Grossberg: as a neural field theory, which combines adaptivity with complex a priori models.

2.14.3 Illusions and A Priori Contents of Vision

In the area of vision, Grossberg discovered neural mechanisms responsible for preattentive processing of visual information in retina, lateral geniculate, and visual cortex. His research, in addition to physical intuition, was guided by analysis of visual illusions. Analysis of illusions was used to deduce a number of specific mechanisms, or a priori models used by the visual system. In particular, Grossberg's group discovered a long sought mechanism of discarding illuminance: how does the visual system achieve a remarkable constancy of perception, independent from variations in illumination. These mechanisms include models used by the visual system to enhance contrast and to complete perceptual boundaries.

Analyzing the concrete content of a priori perceptual models, Grossberg methodologically continues the Kantian approach to elucidating contents of a priori knowledge. Kant's investigation of reason included systematic elucidation of contents of a priori knowledge based on antinomies of reason. Similarly, Grossberg's physics of mind, when analyzing the visual perception, included systematic elucidation of the a priori contents of the vision system based on illusions or antinomies of the visual perception.

Visual measurement is fundamentally limited by being dependent on illuminants, their spectral content (color) and position relative to a scene, it is affected by shading, by not having a direct measurement of depth or range to objects, by not knowing what are the boundaries of objects and what are accidental or shade boundaries, and by the whole set of other ambiguities. Visual perception has to compensate for this uncertainty. This compensation is achieved by using a priori knowledge about what elements of the scene are important and how they should be combined to compensate for the uncertainties. This a priori content, or models of visual perception, form a heterarchical system, that is a system that combines parallel and hierarchical processing. Previous analyses of the visual system assumed existence of independent modules (e.g., for edge detection, segmentation, shape from shading, estimation, object identification); it led to difficulties related to the basic uncertainty of the visual measurement: uncertainties at every processing stage led to more uncertainties at the next stage. Grossberg found that the resolution of these uncertainties is achieved through parallel interaction among several modules, boundary contour system (BCS), feature contour system (FCS), and binocular vision.

The BCS controls the emergence of a 3-D segmentation of a scene. It detects, enhances, and completes boundaries and groups textures. And it performs a matching of the emerging boundaries between the two eyes; this binocular matching process is sensitive to disparity and scale. The outcome of the BCS segmentation process is perceptually invisible. Visible percepts emerge from the FCS. A BCS segmentation regulates the processing of color and brightness by the FCS. This is a hierarchical process controlled by the segmentation hierarchy. Signals within the FCS interact with the BCS signals to control a featural filling-in process, which results in the extraction of color and brightness signals that are relatively unaffected by illumination. These processes lead to visible percepts of color, form, and depth. The overall process is mostly automatic and preattentive; still it influences and may be influenced by attentive object recognition processes.

2.14.4 Motor Coordination and Sensorimotor Control

One difficulty in studying control structures of motor behavior and their coordination with sensory systems is that large numbers of brain regions are utilized to control even relatively simple systems. For example, at least eight different brain regions are involved in a control of saccadic, or ballistics eye movements, a relatively simple type of motor behavior. Studies limited to performance characteristics turned out to be inadequate for discovery of the organization principles coordinating a large number of circuits in a distributed system. Grossberg and co-workers concentrated on developmental and learning problems that have to be solved by a brain system in order to form an adaptive relationship with the environment. This led to a "rapidly expanding understanding of brain mechanisms." Principles of adaptation are fundamental in determining the design of behavioral mechanisms. For example, adaptation requires error correction, but individual neurons are not able to measure the accuracy of movements. Self-correcting mechanisms are embodied in the neural network as a whole.

Consider the problem of learning a coordination between eye–head and hand–arm movement, which is solved by every infant. Each of these two adaptive systems learns a representation, or map, of intended target positions and current position. They are matched in the error-correction process. Coordination between the eye–head and hand–arm systems involves only the target positions. Therefore, learning of intermodal map coordination is gated by intramodal matching. This illustrates the involvement of several circuits in the self-calibrated behavior development. Gated learning is one of the universal principles of sensorimotor control.

Other universal principles include mechanisms of distinguishing self-movement from world movement and two types of control mechanisms associated with continuous and ballistic movements. Distinguishing between self-movement and world movement is needed so that the error-correcting mechanism does not destroy its own correct parametric calibration in response to world motion. Continuous and ballistic motion mechanisms use different approaches to computing target and present positions and to their comparison through time. They also use different approaches for distinguishing between self-movement and world movement.

Among the universal principles is also the need to learn motor synergies, rather than individual muscle commands. Specialized neural networks are employed for coordinated control of multiple muscle groups. These control mechanisms can be characterized as parametric models. In the process of motor behavior, these models are adapted to the need of a specific task by controlling a few of their parameters. Consider, for example, handwriting: having once learned to write, one can write very small or large letters without retraining. And even the individual features of the one's handwriting are preserved. This is related to the fact that in handwriting, as well as in other complicated motions, multimuscle coordination is achieved by changing only few model parameters: the relative motion of muscles remains invariant.

According to Grossberg's models, the main principle of motor learning is the negative feedback and error correction. This is different from the resonant principle of sensory learning, which involves positive feedback. Grossberg suggested that consciousness is related to the resonant state in ART neural networks; this explains why motor learning is usually unconscious.

2.14.5 Emotions and Learning

The field was pioneered by Grossberg, who was far ahead of his time, when in the 1970s he began to study emotions as an integral part of learning and behavior. He has shown that emotional or affective signals act as reinforcers and inhibitors interacting with the formation of sensory representations. Emotional or affective neural circuits are related to basic instincts or drives. Their role is to evaluate the sensory signals as indicators of desirable or undesirable conditions with respect to the satisfaction of the basic needs of the organism. Signals generated by emotional circuits are intended to produce actions that will satisfy the basic instincts. Emotional circuits are of a more ancient origin than logical thinking, they are "more directly" responsible for survival, and therefore they could be expected to effect an omnipresent influence on the thinking process. This fact is well known in psychology, and can be observed by everybody in their own day-to-day decision processes.

Despite these known facts about the role and origin of emotions, the logical rule-based approach of expert and production AI systems does not have evaluative mechanisms similar to emotions. Just the opposite, as we saw in Soar's mechanism of preferences, evaluation is implemented through logical rules. In other words, rule systems attempt to implement emotional mechanisms through logic; this is contrary to the established biological and psychological facts about natural intelligence. And we saw that this evaluation mechanism for competing hypothesis is inefficient; it led to a combinatorial explosion. At the beginning of the AI era, adequate *mathematical* ideas of the role and internal representations of emotions were missing. These mathematical concepts were to a significant extent developed by Grossberg and his co-workers.

ART Fig. 2.14-1b illustrates (at a conceptual level) the existence of neural connections between sensory–cognitive processes and internal drives. These circuits are further elaborated in Fig. 2.14-2. Let us introduce the terminology used for describing this figure. A stimulus is a perceived external event that stimulates a reaction. A designation of "unconditioned stimuli" is used for those stimuli that appeal "directly" to internal drives and do not need to be learned, such as a smell of meat for a hungry dog (it is possible that all or many of "unconditioned" stimuli were learned at an early age). All other stimuli have to be learned, or conditioned by experience, and are called conditioned stimuli (CS). The top of Fig. 2.14-2 shows processes in the F_2 field of an ART circuit. A sensory system faces a continuous stream of signals, some of which resemble previously learned CS. During the process of perception, each CS elicits an internal sensory representation (S_{cs}) in the F_2 field. These representations compete for the short-term memory resources in F_2 (as described in Section 2.14-2 on ART). The activated S_{cs} produces signals to the drive representation D. If these signals are strong enough *and* the drive signal is strong enough, this combination produces a signal from D to S_{cs}, enhancing this particular representation in its competition with other sensory representations. In this way, internal drives produce emotional signals (also called affective or motivational signals), which affect the recognition process in sensory–cognitive circuits.

Note that there are neural cell populations related to particular drives that are separate from cells encoding sensory representations. Repeated occurrences of a particular CS together with activation of a drive representation lead to strengthening synapses of neural connections between the corresponding S_{cs} and D. This is the process of learning

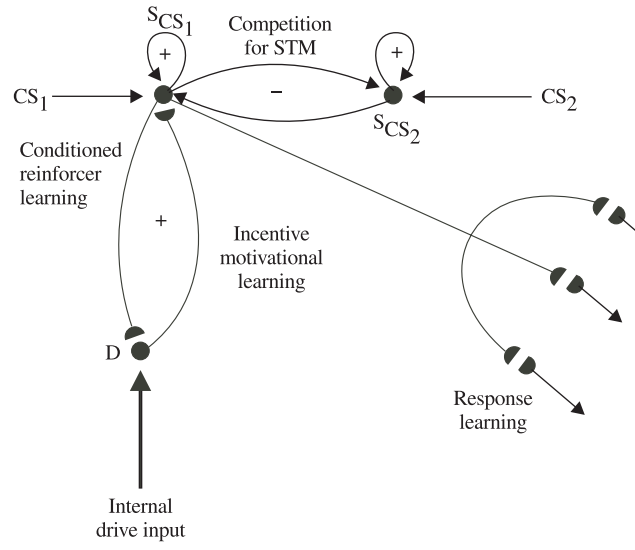


Figure 2.14-2 Emotional signal circuit top-level architecture.

the CS. Strengthening of pathways from S_{cs} to D is called the learning of conditioned reinforcers. Strengthening of pathways from D to S_{cs} is called the learning of motivational incentives. The conditioned S $\rightarrow$ D $\rightarrow$ S pathways shift the attention toward previously reinforced sensory signals, which corresponds to the drives active at a given moment. Thus, a potentially desired object is paid more attention and is preferentially recognized in the continuous stream of signals/stimuli.

A detailed explanation of the temporary dynamics of many conditioning or learning processes requires an additional circuit embedded in the drive representations called a gated dipole. A gated dipole is a neural mechanism of psychological opponent processes, a universal principle shared by perceptive and emotional systems. Dipole fields describe well-known psychological effects. For example, when a red field is presented to an eye, and than suddenly changed to a white field, there is a transient aftereffect: for a short time, the white field is perceived as a little greenish. A similar aftereffect exists in the emotional circuits: a small constant negative stimulus (a shock) naturally generates negative emotions. When the negative stimulus is withdrawn, a positive emotion is generated although "nothing positive" occurred. Grossberg's gated dipole circuit explains this effect due to the interaction of three factors. First, both opponent cells reacting to an onset and offset of a stimulus are constantly in a state of a low-level nonspecific arousal. Second, the opponent cells are mutually inhibiting each other. Third, when a specific stimulus is presented, it excites the corresponding onset cell; while this cell is excited, a gradual depletion of neural transmitters occurs along the excited neural pathways. An offset of the stimulus lead to the offset-cell pathways overtaking the onset ones, whose neural transmitters have been partially depleted. This is a universal principle leading to adaptation to the status quo, so that changes are emphasized over the status quo in perception as well as in emotional states.

Let us summarize in a simplified way the principal difference between the roles of emotional and conceptual–cognitive signals: concepts are representations of the world "as is" (say, "red" or "chair"); emotions are evaluative signals: "good" or "bad." Grossberg has shown that perception, attention, and cognition are not just "logical conceptual processes according to the categories of logic," but are affected by emotions related to the basic instincts. In Chapter 10 we will see that even the most refined and abstract thinking processes, when no basic instinct seems to be involved, are impossible without instincts and emotions. Our analysis will demonstrate that thinking is related to and is impossible without *a basic instinct to learn*. The special types of emotions related to this instinct are responsible for the higher mental abilities, including perception of beauty.

2.14.6 Quantum Neurodynamics

It is well known that our eyes can register a single quantum of light, a photon. However, we know much less about quantum-level mechanisms that might be operational in the neural networks of brain. Are processes at the quantum level important for understanding fundamental psychic phenomena of perception, cognition, memory, affect, and consciousness? An alternative view is that psychic processes occur only as a result of the interaction of very many ions, electrons, and photons. So that all effects specific to quantum physics average out and classical physics is sufficient for understanding of the mind. This latter point of view was prevalent among biologists and neural scientists until recently. But in the late 1980s and 1990s increased attention was drawn to the possibility that quantum processes could be essential for the understanding of the mind. Several factors stimulated this interest. In neurons and synaptic connections there were identified microstructures related to specific mechanisms of information processing in neural networks, and which operate at the quantum level. Quantum laws are seen by some researchers as the key to explaining mysterious properties of the mind. Progress was made in understanding how to build quantum computers.

A quantum system, while transitioning between observable states, may pass through a large number of intermediate states "in parallel." This promises a possibility for the parallel processing of information at a capacity unimaginable for classical computers. This quantum behavior is due to the uncertainty principle, which could be described by saying that a quantum system exists in multiple spatial and temporal locations, and in some sense could "sense its future." Among the first quantum physicists studying biological systems was a founder of quantum mechanics, Schrödinger. His 1944 lectures identified many of the issues that were later studied in a new field of molecular biology that emerged over the next decade. A first attempt to develop quantum field principles of brain organization was made by Umezawa in the late 1960s. This approach was termed quantum brain dynamics (QBD). The essential idea of QBD is that quantum effects are important in the brain on macroscopic scales. Quantum fields and their quanta appear in QBD as collective excitations of a large number of atomic and molecular states, such as the molecular vibrational states in long biomolecules. Quantum fields interact with a classical dynamic system of transmembrane ionic diffusions and macroscopic neural networks. Umezawa and his school developed a mathematical apparatus of quantum field theory applicable to QBD and suggested identification of several quantum field phenomena with psychic processes. Among quantum phenomena emerging in QBD are symmetry breaking, leading

to phase transitions, and emergence of a Goldstone boson, a coherent propagation of a symmetry-restoring wave. Memory formation is explained as a phase transition process, from a higher symmetry to a lower symmetry state. Memory recall is explained due to Goldstone bosons.

Quantum effects on macroscopic scales were observed in laboratory experiments, still there is no evidence that they take place in the brain. Also, from psychological and engineering standpoints, QBD is still in an incipient stage: it has not explained any psychological fact better than theories based on classical physics (like Grossberg's), nor has it generated new mathematical insights into engineering problems. Considering applications of quantum fields to the theory of mind and brain, three directions can be identified: direct use of quantum theory as the first-principles theory of the physical substrate of neural systems, direct use of quantum theory to design quantum computers and quantum neurocomputers, and metaphorical use of quantum theory as a source of mathematical methods for describing processes of classical systems. The role of QBD as the first-principles quantum theory of neural processes was not yet proven. But other approaches seems more promising in this regards, such as analyses of the microstructure of neurons. Also, serious efforts are underway to design "quantum computers." Chapter 8 outlines quantum modeling field theory (QMFT), a description of a quantum device implementing modeling field theory.

2.14.7 Modeling Field Theory

Modeling field theory (MFT) is the main subject of this book. The theory is developed in Chapter 4, and several of its engineering applications are described in Chapters 5 through 7. As a neural architecture, MFT is similar to ART. It makes a next step in the direction of the development of the general mathematical technique combining learning and adaptation with utilization of complicated a priori knowledge (or domain knowledge in engineering applications). A priori knowledge in MFT is utilized in the form of compositional internal models: MFT models are composed of multiple submodels, which can interact or be independent from each other in varying degrees. These submodels correspond to objects or concepts in the outside world or in the "sub-concept-signals" identified at a lower processing level. The learning dynamics of MFT is determined by maximization of similarity between the models and the world.

Philosophically, MFT implements the Aristotelian theory of mind as a priori Forms (models) interacting with matter (sensory signals). Mathematical analysis of MFT, when combined with the semiotical analysis of the nature of signs and symbols, leads to the conclusion that the process of adaptation of an MFT submodel describes a symbol-process. When the mathematical analysis of MFT is combined with the analysis of the role of emotions in cognitive processes and with the Kantian analysis of the structure of mind, it leads to the conclusion that MFT describes a new type of basic instinct or drive, an instinct for increasing the internal knowledge about the outside world, or, in short, an instinct for knowledge. This instinct leads to "higher emotions" serving as a foundation for the ability to perceive beauty. This role of MFT in the total architecture of intelligent systems is described in Chapter 10. It leads to the possibility of the mathematical understanding of beauty as a property of very complex adaptive systems, a mechanism of adaptation related to the overall system goal, which might not be entirely specified a priori.

2.15 Intelligence, Learning, and Computability

2.15.1 Computability: Turing vs. Physics

In 1900, Hilbert formulated 23 fundamental mathematical problems, the purpose of which was to complete the axiomatization of mathematics and to define forever the rules of mathematical procedures. Hilbert's tenth problem, Entscheidungsproblem, was to answer the question if, in principle, there can exist a universal mathematical procedure for solving any mathematical problem (within a definite but sufficiently reach domain). The answer was found by Turing, who proved that a universal mathematical procedure does not exist. In search of the proof, Turing formulated the concept of algorithmic computability as an existence of a solution to a mathematical problem that can be computed in a finite (unknown beforehand) number of steps. A step is one of several simple predetermined operations, such as addition, or a selection of the next operation depending on the result of the previous one (e.g., a computer language statement).

When analyzing concrete algorithms in connection with the existence of procedures of the MHT type, which, although finite, may require an unphysically large number of steps, the Turing formulation of algorithmic computability is insufficient. A physical algorithmic computability, accounting for computational complexity of algorithms, can be formulated as a bound on how fast the number of steps or operations grows with the growth in complexity of the problem. To define the concept of physical computability, let us introduce a parameter or a set of parameters, D, characterizing the problem complexity; D might include the dimensionality of the classification space, the number of classes, the number of measurements, etc. If the number of required operations, $N > D^n$, for any n [nonpolynomial complexity, such as $\exp(D)$], the algorithm is of a physically noncomputable type. If a number of required operations, $N < D\cdot$ constant for some finite value of the constant, then an algorithm is physically computable; but if in addition there exist a parallel algorithm formulation that permits using P parallel processors to solve the problem in $N < D/P$ steps, then the algorithm is physically computable in real time. That is, by increasing the number of processors with increases in the complexity of the problem, a solution can be achieved for the problem of any complexity within a constant (nonincreasing with the problem complexity) time interval.

There exists a field of mathematics studying computational complexity of algorithms. A fundamental result obtained in this field states that there are several classes of computational complexity of problems. Some problems can be solved by algorithms of polynomial computational complexity, while others are inherently nonpolynomial (e.g., exponential or combinatorial) in complexity. In other words, there are nonpolynomial problems for which no polynomial algorithm can be found. These results are now being questioned by a theoretical possibility of quantum computing. A quantum system may exist at once in a superposition of multiple states. For example, electron motion can be described as a concurrent "evaluation" of a large number of combinations of various trajectories. But an electron moves in finite time; it does not need combinatorial time to figure out how to move. If a quantum system can be devised, where electron motions are used to "compute" various parts of a problem along each trajectory segment, such a system would "violate" the difference in principle between polynomial and nonpolynomial problems. Whether a general purpose quantum computer with the desired property can be built is still a subject

of some controversy. But the fact that every quantum system is a "special-purpose" device that efficiently computes its own states seems to violate the basic tenets of the existing computational complexity theory.

2.15.2 Computational Methods of Intelligence: Summary

Let us summarize the discussion of the mathematical methods of describing intelligence reviewed in this chapter. A physical theory of mind ought to satisfy first, apriority (or an ability to utilize complicated knowledge); second, adaptivity (or an ability to learn from a limited number of examples in a changing world), and third, physical computability in real time (or an ability to solve the problems of sufficient complexity within the limited time interval). Classical mathematical methods, as we have seen, do not satisfy these requirements and are not appropriate as a foundation for physically acceptable concepts of intellect. A number of new computational methods emerged during the 1980s and 1990s that attempt to combine apriority, adaptivity, and physical computability. The most promising among them include neural fields, hierarchical organization, evolutionary computing, and fuzzy logic. The remaining part of this book is devoted to the development of the modeling field theory of mind that promises to fulfill a need for the physical theory. We will also discuss its relationships to the other emerging computational concepts and ways of combining their most promising aspects.

NOTES

1. Let us explain again the origin of the term *combinatorial*. In many algorithms it is needed to consider combinations of N points by M groups (for M models). The number of these combinations is $N^M = \exp(N \ln M)$. For medium-size N and M this number is very large and it grows fast with N and M. Fast-growing functions of this type are called combinatorial or exponential.
2. Throughout this book, when referring to Aristotelian logic, I always mean the contemporary formalized understanding of the logic developed by mathematicians in the eighteenth and nineteenth centuries from Boole to Russell. Aristotle considered logical statements to be linguistic sentences. He was aware of the uncertainties inherent in language and in every statement. In *Metaphysics*, he emphasized that it is necessary to use common sense to appreciate these uncertainties in order to use logic properly. In this way, Aristotelian understanding was closer to the concepts of fuzzy logic and granulation developed by Zadeh. However, the uncertainty inherent in the language goes against the "law of contradiction (excluded third)." Aristotle did not reconcile the uncertainties with his law of contradiction. This reconciliation was achieved by formal logicians from Boole to Russell, who eliminated linguistic uncertainties. Gödel proved that this "exactness" is inconsistent. Zadeh achieved the reconciliation by eliminating the law of contradiction.
3. See the discussion of the segmentation and data association problem in Chapter 1, Section 1.3.3.
4. This is not an exaggeration. Human reaction time is on the order of 500 ms, while neuron firing time interval is about 5 ms. Therefore, a typical human perception involves about 500 ms/5 ms = 100 sequential steps. Consider 100 sequential decisions made by using one of R available rules at each steps (say, $R \sim 10$ to 100). All possible combinations N_c of applying these rules is $N_c = R^{100}$, or $\lg N_c = 100 \lg R \sim 100$ to 200. The number of all elementary particles N_p in the universe is on the order of $\lg N_p \sim 108$, and the number of all elementary particle interactions in the whole history of the universe, N_{ei}, is on the order of $\lg N_{ei} \sim 150$.

5. Mathematical theory of symbol is discussed in Sections 2.12 and 10.3.4.
6. See note 4.
7. Bayesian networks are directed acyclic graphs.
8. See note 2.
9. See note 5.
10. This view of Marvin Minsky is shared by many of his colleagues, whose long and productive careers, nevertheless, did not lead to finding "elements of intelligence." In the fall of 1997 a pioneer of the area of pattern recognition, Laveen Kanal, retired from his professorship at the University of Maryland. His colleagues organized in his honor a workshop, "Intelligent Systems." The workshop's opening talk was given by A. Rosenfeld, an influential scientist and one of the founders of the area of pattern recognition. He pronounced that in the area of artificial intelligence all the main discoveries have already been made, and what remains is just to apply them using more and more powerful computers. This reminds me of another story. When Max Plank, a discoverer of the quantum nature of light, was still in college (in the 1880s), his professor told him essentially the same things about physics.
11. Researchers in the field of neural networks have long argued for the inadequacy of rule systems to explain learning. But Minsky should be credited with being the first among "insiders" and founders of rule-based approach to emphasize that it is fundamentally limited by an inability to learn.
12. Albus and Meystel's terminology with regards to differentiating adaptation as a simple behavioral mechanism vs. learning as a more complicated increase of intelligence is not universally agreed on. For example, Holland uses the term *complex adaptive systems*, which combines both types of functioning.
13. In the unabridged Webster's dictionary, the first meaning of symbol is "a creed of faith,"—something that has a spiritual power. Jung uses "symbol" in a similar way: it is a psychological process in which a new meaning is created. Pribram calls signs within the brain the less adaptive signals and symbols within the brain the more adaptive signals.
14. Thus, learning in ART occurs only in a resonant state corresponding to a high degree match between an input signal and previously stored models. This sets ART apart from self-organizing feature maps (Malsburg, 1973; Grossberg, 1976; Kohonen, 1984). In self-organizing feature maps, learning occurs in all connections whose F2 nodes (feature vector components) win the competition. This type of learning may lead to forgetting (due to relearning) of the previously learned patterns. ART avoids this problem, because the resonance requires a high degree match (Grossberg and Merrill 1992).
15. In this paragraph, as well as in most of this book, "a priori" refers to the state of the internal model prior to the current learning experience. It contrasts a classical usage of this term, which refers to "God-given" unmodifiable contents that transcend all experience. The understanding of "a priori" as nonmodifiable by any experience ascends to Plato's Eidos; it caused difficulties for the philosophical understanding of mind as well as for the mathematical methods of AI, in which a priori representations are not sufficiently adaptive (e.g., crisp rules).

BIBLIOGRAPHICAL NOTES

This section contains additional bibliographical information that was not explicitly referenced in the text.

Abductive reasoning (Bhatnagar and Kanal, 1993).
Additive neural network (McCulloch and Pitts, 1943, see 1948; Grossberg, 1968; Hopfield, 1982).

Artificial intelligence history (Newell, 1987; Gardner, 1987; Crevier, 1993; Franklin, 1995).

Classification space partitioning (Perlovsky, 1994a).

Chomsky, on similarity between rule systems and Plato's Ideas (Chomsky, 1972); principles and parameters (Chomsky, 1981).

Chomskyan linguistics, review (Botha, 1991); on the language faculty as a system of grammatical rules (Moravscik and Wirth, 1980; Berwick, 1982).

Combinatorial complexity (Minsky and Papert, 1988; Winston, 1984; Negahdaripour and Jain, 1991; Grimson and Huttenlocher, 1991).

Combinatorial complexity: training complexity (Bellman, 1961; Perlovsky, 1994a); complexity vs. a priori knowledge (Perlovsky, 1991, 1994a; Girosi et al., 1995); rule and model-based systems (Winston, 1984).

Competitive learning networks (Grossberg, 1970; Kohonen, 1984). Other neural networks based on the nearest neighbor concept (Reilly et al., 1987; Specht, 1990; Fukushima, 1980; Poggio, 1988; Carpenter et al., 1991a,b; 1992).

Computer's thinking (Searle, 1980; Penrose, 1994); intelligence as varying degrees of thinking (Minsky, 1985; Franklin, 1995).

Discriminating surfaces concept (Duda and Fossum, 1966; Ho and Agrawala, 1968; Specht, 1967; Nilsson, 1967).

Evolutionary computation, genetic algorithms, and complex adaptive systems (Holland, 1992, 1995; Fogel, 1995).

Feedforward neural networks (Werbos, 1974; Parker, 1982; Rumelhart et al., 1986); their analyses (Moore and Poggio, 1988; Carpenter, 1989; Longstaff and Cross, 1987; Lang et al., 1990); combinatorial nature of training requirements (Minsky and Papert, 1988; Blum and Rivest, 1992); overtraining (Lang et al., 1990).

Grossberg and co-workers: on ART (Grossberg, 1976, 1980a; Carpenter and Grossberg, 1987; Carpenter et al., 1991a,b, 1992); on distributed ART (Carpenter, 1994, 1996, 1997; Carpenter et al., 1998); on ART applications (Carpenter et al., 1997); on visual perception (Grossberg, 1982, 1988a); on sensorimotor control (Grossberg and Kuperstein, 1989); on emotional and affective neural circuitry (Grossberg, 1972a,b; Grossberg and Schmajuk, 1987; Grossberg and Gutowski, 1987; Grossberg and Levine, 1987; Grossberg and Merrill, 1992).

Hierarchical organization: Meystel (1988), NIST model (Albus, 1991; Meystel, 1996).

JDL model described (Hall, 1997; Hall and Linn, 1990; Hall et al., 1991). Further reading on various aspects of sensor fusion related to JDL (Llinas and Waltz, 1990; Klein, 1993; Bar-Shalom and Li, 1995).

Kant on beauty (1790).

Mathematical algorithms of many neural paradigms and their relationships to the NNC have been analyzed by Simpson (1990).

McCulloch's work on a material basis for concepts or ideas of objects independent of their apparent size (Pitts and McCulloch, 1947).

Minsky's work on frame theory (Minsky, 1968); rule systems are inadequate for adaptation and learning (Minsky, 1968, 1975, 1985).

Model-Based Vision, MBV (Nevatia and Binford, 1977; Brooks, 1983; Winston, 1984; Grimson and Lozano-Perez, 1984; Chen and Dyer, 1986; Michalski et al., 1986; Lamdan and Wolfson, 1988; Negahdaripour and Jain, 1991; Bonnisone et al., 1991; Segre, 1992; Keshavan et al., 1993; Califano and Mohan, 1994; indexing (Lamdan and Wolfson, 1988; Califano and Mohan, 1994).

Model-based approaches (Nevatia and Binford, 1977; Brooks, 1983; Winston, 1984; Grimson and Lozano-Perez, 1984; Chen and Dyer, 1986; Michalski et al., 1986; Lamdan and Wolfson, 1988; Negahdaripour and Jain, 1991; Bonnisone et al., 1991; Grimson and Huttenlocher, 1991; Segre, 1992; Keshavan et al., 1993; Califano and Mohan, 1994).

Neural paradigms reviewed (Simpson, 1990; Haykin, 1998; Bishop, 1995; Cherkassky and Mulier, 1998). The latter analyzes typical neural paradigms from the Statistical Learning Theory perspective.

Parallel realization of systems of rules (Rumelhart and McClelland, 1986).

Parametric vs. nonparametric techniques (Tukey, 1960).

Pattern recognition algorithms (Nilsson, 1965; Fukunaga, 1972; Duda and Hart, 1973; Watanabe, 1985).

Quantum effects in biological systems (Schrödinger, 1944; Stuart et al., 1978; Hameroff, 1987; Penrose, 1989, 1994).

Quantum field theory approaches to neural networks (Pribram, 1993). This volume contains papers presented at the First Appalachian Conference on Neurodynamics. My review of QBD followed the paper by Jibu and Yasue (1993) from this volume. Also, a "metaphorical" description for some utilization of QFT methods is borrowed from Werbos (ibid.); quantum modeling field theory (Perlovsky, 1997c,f).

Reviews of computational principles (Minsky and Papert, 1988; Perlovsky, 1991a; 1994a; Franklin, 1994; Bishop, 1995; Cherkassky and Mulier, 1998).

Rule systems (Minsky, 1968a,b; Winston, 1984); learning (Winston, 1984); uncertainties and combinatorial complexity (Chapman, 1987; Maes, 1991; Franklin, 1995).

Rule-based agents (Maes, 1991).

Semiotics, founders (Peirce, 1935–66; Morris, 1971).

Soar (Laird, 1981).

Statistical Learning Theory, SLT (Vapnik, 1995); its more accessible exposition can be found in Cherkassky and Mulier (1998); it also contains reviews and analyses of neural networks in terms of SLT (Cherkassky and Mulier, 1998).

Turing (1937).

Wiener, on inadequacy of simple models, because of insufficient data for learning (Wiener, 1948).

Zadeh works on fuzzy logic and granularity (Zadeh, 1962, 1965, 1997).

100-step rule mentioned in Note 4 was described by Franklin (1995) and attributed to Feldman (unpublished).

PROBLEMS

2.5–1 Compare the MHT description in Section 2.5.3 with the intelligent tracker description in Section 1.1.1.4.

2.6–1 Establish a correspondence between the definitions and equations in Section 2.6.2 and Fig. 2.6-1.

2.7–1 Prove that the nearest neighbor training leads to $\min R_{\text{emp}}(p) = 0$.
Hint: Define parameters $p^* = \{y_n\}$; predictive models are defined as follows, for each $\mathbf{x}$, select the closest $\mathbf{x}_{n'}$ among the training set $\{\mathbf{x}_n, y_n\}$ and $M(\mathbf{x}, p^*) = y_{n'}$.

2.12–1 Compare contents of this section with the intelligent tracker description in Section 1.1.4.

2.13–1 Verify that $\bar{g}_i = r_i$, and $c_i = \overline{(g_i - r_i)^2} = r_i(1 - r_i)$.
Hint: Use the definition of the average: $\bar{g}_i =$ [sum over population of (1, for $g_i = 1$) + sum of (0, for $g_i = 0$)]$/N$, were N is the total number of agents in the population; denote N_{i0} and N_{i1} the numbers of agents with $g_i = 0$ and $g_i = 1$. Then, $\bar{g}_i = N_{i1}/N$; denote this r_i. Now, in a similar way compute $c_i =$ {sum over population of $[(1 - r_i)^2$, for $g_i = 1] +$ sum of $[(0 - r_i)^2$, for $g_i = 0]\}/N$.

MATHEMATICAL VERSUS METAPHYSICAL CONCEPTS OF MIND

> The danger is in the fateful "fear of metaphysics" . . . Therefore, I was especially pleased to find out . . . that, at the end, we can not do without metaphysics.
>
> —EINSTEIN

Mathematical modeling of intellect, the beginning of which dates to the 1940s and is contemporaneous with the computer age, is undergoing today an explosion brought about by new ideas in the theory of neural networks—a theory modeling the neural architecture of the brain with the purpose of explaining the mind. Ever since antiquity philosophers, theologists, and scientists have attempted to uncover the mystery of mind. Could the old philosophical discussions help with contemporary mathematical development? Is contemporary neural network research related to eternal philosophical questions? Will it resolve the mystery of mind like Newtonian physics resolved the mystery of matter?

This chapter traces continuous connections of concepts of mind, in seemingly so completely different thinkers, separated by time, culture, and geography, as pagan philosophers of ancient Greece, theologians of monotheistic religions, and scientists of today. Various current directions in the neural network theory are connected, despite enormous gaps in time, to various concepts of mind, suggested more than 2000 years ago. Inquiring how the intellect combines apriority (an ability to utilize a priori, preexperiential knowledge of eternal truths) and adaptivity (to an ever-changing world), I will attempt to demonstrate that one and the same problem of how the intellect is being determined by both these factors has ever been turning at the center of philosophical, theological, and scientific debates on the nature of mind. The enormous treasure of the world's knowledge intrigues me with a possibility of combining remote points of knowledge, of determining main landmarks in the approaches to pressing problems, and of understanding the significance of various algorithms and paradigms. Connections between philosophy and mathematics established here clarify relationships of the mathematical concepts to mind and help in defining promising research directions.

3.1 Prolegomenon: Plato, Antisthenes, and Artificial Intelligence

In the reflections of ancient philosophers on the nature of the pure spirit, despite the enormous time gap, I am finding concepts of the intellect that are directly related to contemporary mathematical ideas on the nature of mind. A picture of our contemporary world is not at all new, and an understanding of the physics and mathematics of mind that is emerging today at times happens to be nearer to certain philosophical systems and concepts created in the course of more than two millennia, than to the concepts of pattern recognition and artificial intelligence developed only a few years ago. And between Plato and Minsky, Aristotle and Grossberg, Aquinas and Jung, Kant and Chomsky there is such a close connection, as if some ideas and concepts belong to each other, and time is only a cloud interfering with our seeing this cobelonging.

This chapter traces continuous connections of concepts of mind through the time, cultures, and geography to contemporary mathematical algorithms. The concepts, hypotheses, and controversies of the past considered in terms of today's research issues open the tremendous wealth of accumulated knowledge hidden behind a veil of changed meanings of words and historical events. I will attempt to utilize this rediscovered knowledge to analyze the perplexing issues of today's mathematical research of intellect, to outline future directions of neural network research, and possibly to approach the comprehension of the mystery of the nature of mind.

The very first question about the intellect is: How is it possible at all? Plato's answer was that the concepts of mind must have been put into a human being a priori, that is before the existence of the individual human being. This philosophical concept was named "the realism of ideas." The original, Platonian use of the word *realism* refers to the reality of a priori ideas as opposed to the nonreality of experience. In the course of millennia the word *realism* has been used in different ways, acquiring sometimes the opposite meaning, denoting thinking grounded in experience. In this book, *realism* always has the original Platonian meaning of the realism of a priori ideas.

Another answer to the first question about intellect was given by Antisthenes. Like Plato, Antisthenes was Socrates' pupil. But his answer was contrary to the Platonian one. According to Antisthenes, Ideas have no real existence independent from individual objects of experience. Ideas are just words designating classes of similar objects. The philosophical concept descending from Antisthenes was named *nominalism* (from Latin *nomina* = name). A schism between nominalists and realists determined philosophical debates throughout the Middle Ages and continues to exert a profound effect on mathematical concepts of intelligence in our time.

Plato's principle of apriority was used by Minsky as a basis for creating computer artificial intelligence. For a computer to operate and make decisions in a complicated environment, concluded Minsky, knowledge ought to be placed into the computer a priori (Minsky, 1968a). In Minsky's method, named the expert or rule systems, a system of logical rules is put into a computer, containing all possible situations (for example, all possible readings of sensors of a particular device or system), and expert decisions or rules of what is to be done in each particular situation. This method, which I call the Plato–Minsky approach,[1] became the foundation for many practical applications of computers from factory floors to space shuttles.

Solving the very first problem of intelligence: How is it possible?—the Plato–Minsky approach does not explain an important aspect of mind—an ability to learn and to adapt, leaving unanswered the second question about the intelligence: How is learning possible? In Platonian theory, learning is addressed as an afterthought. The second question about mind was first addressed by Antisthenes: according to nominalism, we learn concepts and ideas by "classifying" objects according to their similarities. This learning process leads to general ideas and concepts. In a different way, this question was addressed by Aristotle, who combined both properties of mind, apriority and adaptivity.

Although Minsky emphasized that his method does not solve the problem of learning, attempts to add learning to Minsky's artificial intelligence continued in various fields modeling the mind, including linguistics and pattern recognition. In linguistics, Chomsky proposed building a self-learning system that could learn a language similarly to a human, using a mathematics of logical rule systems. In Chomsky's approach, the learning of a language is based on a language faculty, which is a genetically inherited component of the mind, containing an a priori knowledge of language. A direction in linguistics, named the Chomskyan Revolution, was about recognizing the two questions about the intellect (first, how is intellect possible? and second, how is learning possible?) as the center of linguistic inquiry and of a mathematical theory of mind.

A different direction of research into the mathematical concepts of the intellect, founded by McCulloch, is based on modeling the neural structure of the brain. Many cognitive scientists following this approach have been convinced that the neural brain processes are of a different nature than symbolic rule systems. And that the problem of learning cannot be solved by improving the Plato–Minsky method, which is founded on static, nonadaptive logical rules similar to the *ideas* of Plato. Aristotle was the first to point out the absence of learning in Plato's theory and to begin the quest for combining apriority and adaptivity. The ways in which the intellect combines apriority with adaptivity has remained at the center of philosophical, theological, and mathematical debates on the nature of mind.

And which way will the pendulum of contemporary science swing?—toward adaptivity or apriority? Where will the mystery of mind be hidden—in the infinity of Jung's subject or Kant's object? How will this be reflected in tomorrow's mathematical concepts of the intellect, and what will the physical picture of the mind be?

3.2 Learning from Aristotle to Maimonides

3.2.1 The Controversy of Aristotle

Aristotle set on to overcome the limitation of Plato's theory of mind related to the static, nonadaptive nature of the Ideas or concepts of mind. In Aristotelian theory of mind, the a priori contents of mind are not the concepts, but Forms having dynamic, adaptive nature. Forms are, as it were, a bridge between the transcendental world of the a priori spirit and the world of everyday experience. Forms-as-potentialities belong to the a priori content of mind, whereas Forms-as-actualities come immediately close to the world of experience. Mind's adaptation occurs in the dynamic process of meeting between the a priori Form and matter. Individual experience is *formed* by this process. Aristotle illustrated his theory with an example of learning in an individual person as an actualization of potentiality of the a priori-present Form. And he placed these Forms-as-potentialities in the intellectual part of the psyche.

Further, Aristotle reduces the end cause to the formal cause. This can be understood in contemporary terms by relating the end cause to the intentional psychological states and relating the formal cause to the a priori content of mind (that is to Forms). The intentional states present a mystery for the traditional artificial intelligence and their explanation is considered to be among the most important challenges to the contemporary science of mind. Reducing the end cause to the formal cause means that the intentional states of mind should be explained through the a priori contents of mind. The mathematical apparatus necessary to describe this process is developed throughout this book. Aristotelian concept of Forms rejects the self-sufficiency of Plato's Ideas for the more complicated dynamic concept of the mind combining apriority with adaptivity.

Contrary to this dynamic concept of Forms, which is the foundation of Aristotelian metaphysics, Aristotelian physics is based on the eternal invariant principles,[2] unchanging circular motion of the celestial spheres—in essence a static philosophical concept, in which (while matter is moving) the first mover, that is God, does not change in time. The theory of static eternal rotation has led him to postulate a complicated mechanism of the celestial spheres, by which the first cause God affects the material world. The mechanism of the celestial spheres, later developed into the theory of emanation by neoplatonics, was the most objectionable part of the Aristotelian system to many scientists, including Newton.

Discovering the laws of spirit and the laws of matter, Aristotle outlined the contours of the future divorce. However, he embraced neither monotheism nor dualism. The Aristotelian system, a pinnacle of man's thought, still suffered from the half-hearted dualism of the heathen religion of ancient Greece. Aristotle, as it were, sanctioned the incomplete dualism between the causality and eternity of physical sciences on the one hand and the adaptivity of the intellect on the other. This dualism has pervaded philosophical thinking ever since. Although the incompleteness of the dualism has been rejected in one way by monotheism and in another way by Descartes, the nature of the Aristotelian quest is related to the same problem we are facing today: the inadequacy of our rational and mathematical concepts for understanding the intellect as a unified system.

One part of the Aristotelian heritage, the one he himself was proud of and cherished the most, is Aristotelian logic. Aristotelian logic serves as a foundation for many of our algorithms. Concepts of Aristotelian logic are unchanging eternal truths. Similar to Aristotelian physics, it was halfway divorced from its spiritual basis. Created to provide the mind's law, the static nature of Aristotelian logic contradicted the Aristotelian theory of Form. In the ensuing 2300-year debate about the apriority and adaptivity of mind, Aristotelian logic supported Platonian Ideas, not Aristotelian Forms. This contradiction was a subject of detailed discussions in the previous chapter. During the thousand years after Aristotle, the differences between him and Plato were minimized[3] by their followers in order to battle the nominalistic philosophical idea about the nature of mind. Since antiquity, throughout the Middle Ages, and until today, the two properties of mind, its apriority and adaptivity, led to philosophical and scientific schisms.

3.2.2 Finite Angels of Maimonides

The problem of the origin and nature of ideas, or universal concepts, continued to create controversies during antiquity as well as in the Middle Ages. The philosophical controversy related to the adaptivity of the fundamental concepts, first recognized by Aristotle with

regard to Plato's ideas and unresolved in Aristotelian physics, remained an unresolved riddle along the way toward a unified physical and metaphysical view of the world, a controversy manifested in a disconnection between the infinite, invariant Divinity and the changeable, finite Nature. Creation of the unified system on the foundation of Plato's and Aristotle's teachings, undertaken by the philosophers of the Neoplatonic school and especially by Plotinus, demanded an analysis that gradually revealed a fundamental incongruity of the philosophical concepts of unity and adaptivity. In the polarization between monotheistic and Greek theology, the highest principles of Greek philosophy became ever more transcendent and removed from the material world. To restore the relevancy of the transcendental principles for the changing world, additional intermediary disembodied beings were needed to bring the spiritual emanation of God into the material world of the Greek cosmology. But the nature of interaction between the a priori spirit and the dynamic material world could not be understood. For many pagan philosophers of antiquity and the first centuries AD, the static nature of the first principles related to the static notion of divinity was not only obvious, but the only possible nature of ideal universal concepts. In addition, it was perceived to be blessed by the authority of Aristotle. The dynamic, adaptive aspect of Aristotelian theory of Forms was not noticed. The difficulty of comprehending the dynamic aspect of the universal concepts had fundamental theological roots. An idea of Creation that is fundamental to monotheistic religions was perceived as philosophically unsound, anthropomorphic, and naively primitive. To entertain adaptivity of the universal concepts of mind for a philosopher of antiquity was even more difficult than for a physicist in the eighteenth century to consider that Newton's laws will change with time. However, as a result of analysis revealing the problems of the philosophical system previously outlined, the monotheistic tradition appeared less naive to a philosophical mind. This prepared the way for the nexus between theology and philosophy that flourished throughout the Middle Ages.

The problem of adaptivity of the fundamental concepts in medieval theology—in essence, the same problem that Aristotle and Plato were solving in the area of the pure spirit and that we are solving today in the area of the mathematics of intellect—is how to reconcile the absolute nature of ideas with adaptivity and learning. A link of contemporary research in mathematics of the intellect to philosophical and theological debates of the past is possible because the essence of God for the philosophers and philosophical theologists is intellect. In the sixth century BC, Xenophanes understood God governing the world, like a thought governing the body. Therefore, the discernment of philosophers into the nature of the pure spirit as well as the discernment of theologists into the nature of God are immediately important to the mathematical theory of intellect that is being built today. A schism between the rational basis of philosophy and irrational basis of religion has often led to a conflict that is not yet quite resolved today and that counters the nexus of the rational and irrational understanding of the human nature in the unified science of intellect. A touching of theology and philosophy during the Middle Ages was prepared as much by the analysis of philosophical contradictions previously discussed as by the analysis of contradictions within the monotheistic theologies and by the development of theological philosophy.

The first systematic philosophy of the monotheistic concept of creation was, possibly, the Islamic dialectical theology of Kalam. A controversy between the infinite nature and the immediacy of the Deity was formulated and addressed in Kalam as a problem of the attributes of God: divine transcendence vs. immanence. (In today's language, this is the

same problem of apriority vs. adaptivity.[4]) Building on this tradition, Avicenna (ibn Sina) in the eleventh century bridged the philosophical tradition of Plato, Aristotle, and Plotinus with that of monotheism.

Intellect, on the one hand, is an essence of God, and on the other hand, according to Avicenna, is a rational faculty. So, the existence of God is connected to the existence of the self and is proved through the thought (a forerunner of Descartes' *cogito ergo sum*). Thinking that knowledge comes from within, from intuition, in a rational way, Avicenna introduced the concept of estimative intellectual faculty (a forerunner of the Kantian ability for judgment and Jungian concept of the rational feeling function). The intellectual faculty along with imagination is responsible for acquisition of knowledge: knowledge originates from the meeting between the Universal Forms of the intellect and the particular forms of matter. This consideration of how the intellect combines a priori knowledge and adaptivity has further developed the Aristotelian concept of Form.

In the Avicennian concept of material intelligence, resembling the primary matter, we find a forerunner of the Jungian collective unconscious. Even more so, Avicenna considered common sense to be a memory rather than a rational faculty. In his analysis of *being*, he differentiated *being* as essence and being as existence, a concept that in contemporary language is similar to differentiating between meaning and representation, or between the semantic and syntactic contents, an essential feature of contemporary analysis of language and intellect. The teachings of Avicenna are perceived today as strikingly contemporary indeed.

In the next century, mysterious questions of the human mind were studied by Maimonides, who undertook to combine philosophy and theology. For Maimonides, absolute intelligence had a personal nature. A monotheistic God is relevant every moment in the changing world; in today's scientific language, this property of modeling intellect is called adaptivity.[5] To restore the philosophical possibility of adaptive intellect he felt it necessary to reject the Aristotelian physics with its eternity of celestial spheres. Difficulties along this path involved the authority and popularity of the great philosopher. On the basis of Aristotelian teachings, during the past centuries, the efforts of many philosophers built a grandiose system of the unified world view that included Aristotelian physics along with the eternal emanation of spirit, bringing the intellect of God into the material world.

Maimonides disentangled the Aristotelian concept of intellect, as emanation of Form, the eternity of Aristotelian physics, the self-sufficiency of Plato's ideas, and the eternal emanation of Plotinus. By maintaining that emanation of intellect is consistent with the finiteness of the world, Maimonides reconciled Aristotelian physics with Aristotelian metaphysics, and resolved the difficulties of Neoplatonic philosophy involving infiniteness and adaptivity. Maimonides found the much sought philosophical nexus between the adaptivity of the intellect and the infinity of its a priori nature. This connection of the a priori spirit and its adaptive manifestation in the material world was achieved through the analysis of the finite nature of the Aristotelian disembodied intellects or angels. The conceptual step toward understanding mind consisted in suggesting the finiteness of certain spiritual entities. Spirit and matter were brought closer together. Making such a step without sacrificing either matter or spirit is among the toughest challenges that we still face. Finite angels of Maimonides that mediate between the absolute and the particular are the forerunners of efferent signals in Grossberg's neural networks, mediating between the a priori neural structure and the individual objects of perception.

3.2.3 Nexus of Aquinas

A confrontation between philosophical schools and religions that lingered on for millennia yielded to a convergence of philosophy and religion in Christian scholastic tradition. Philosophy and religion came to proximity in the study of the nature of universal concepts of the intellect.[6] The influence of Aristotelian physics on medieval scholasts led many to accept continuous emanation of the intellect into the material world, despite its contradiction of the concept of Creation. A reconciliation of this difficulty was undertaken by Aquinas, who continued the tradition of Avicenna and Maimonides.

According to Aquinas, sensation and intellection are structured in a similar way, in that both have quidditive and existential aspects. Here, quiddity or "somethingness" refers to *what in particular* is sensed or intellected, and the existential aspect refers to the inner or universal content of the experience. Today, Grossberg assigns these aspects of sensation and intellection to afferent (coming from outside) and efferent (coming from inside) signals. And, like Aquinas, he considers this to be a universal law of mind. Thus, the concepts of intellect formulated in the thirteenth century by Aquinas are close to the views of a contemporary philosopher and originator of neural network theory Grossberg, and are directly relevant to current neural network research.

Are there uniform laws of mind that operate at every level from simple perception to cognition of complicated concepts? In the previous chapter we discussed that this is one of the pressing question of contemporary mathematical research into the nature of intellect. Many researchers do not believe that this is possible. For some, perception seems too simple to require the same laws as cognition. For others, cognition seems too personal to ever be described mathematically. Still others think that there is nothing special about either; neither perception nor cognition requires specific sophisticated mechanisms; they believe that the nature of mind is explained by a large number of nonintelligent specialized agents, not obeying any universal law. Together with Aquinas and Grossberg, we have argued that a two-level structure that relates internal representations to input data is a universal law of mind.

3.3 Heresy of Occam and Scientific Method

3.3.1 Cynics, Occam, and Empiricism

Near the end of the period in which scholastic thought thrived, and foreboding the scientific era, the views of Aquinas and other followers of the realistic philosophy of Plato and Aristotle, analyzing the a priori nature of intellect, started losing popularity and were criticized by philosophers perceiving a limiting, determining influence of the concept of apriority. Among the most notable critics of the realism was Occam. He lived in the fourteenth century and is considered one of the last great medieval scholastic thinkers. He adhered to the philosophy of *nominalism* that opposes realism. Nominalism was founded by Antisthenes, a pupil of Socrates, a contemporary of Plato, and the founder of the Cynic school of philosophy. Nominalism, which considers ideas to be just names for classes or collections of similar empirical facts, was extensively developed during the scholastic era in opposition to realism. The opposition of realism and nominalism is not an accidental dispute among scholastic philosophers, but a fundamental issue of the entire philosophy

related to the origin of knowledge, and, in particular, to the origin of the universal concepts of mind and their relationships to individual objects of the empirical world. The schism between realism and nominalism runs from ancient Greece through the Middle Ages and through the current mathematical concepts of intellect. This latter connection, however, is not appreciated within the scientific community; as a result, progress toward understanding the fundamental limitations of various mathematical concepts is slow. Connecting debates in mathematical intelligence to the philosophical ones is essential to understanding the mind.

Occam came to believe that only particular experiences have real existence and that general concepts (universals) are names for similar types of experiences. Assigning universals to the domain of linguistics, he then argued that linguistic and mental phenomena are of an individual nature and should not be considered a reality, and he considered thinking to be ontologically prior to language. It is interesting to note that a similar combination of views on thinking and language is found in the twentieth century among many psychologists of the behavioristic school. Analyzing the empirical, experiential origin of knowledge, Occam developed the basis for the coming philosophy of empiricism. His work indicated (or initiated?) a shift of interest *away* from mental processes, *away* from the question of the possibility of the intellect, and toward an objectified method of inquiry, which later became associated inseparably with the scientific method. Occam's doubts about eternal apriority got him in trouble with the Church: he was arrested and then lived in exile. But the nominalistic concept that knowledge originates in experience seems so obvious that many scientists accept it without doubt.

Time has obscured the influence of Occam on the development of the scientific method, and his name is hidden behind the figures of great philosophers and scientists that came after him. However, despite the realism of Descartes, Leibnitz, and Newton, nominalism as the forerunner of contemporary scientific thinking continues to pervade scientific attitudes of today. One of the reasons for the influence of nominalism is the unbreakable tie between the scientific method and objectification of the subject of inquiry.[7] In physics, the theoretical tradition of Newton's realism counterbalanced the influence of nominalism, but in the area of empirical sciences, such as psychology, the reality of facts seemed more significant than the reality of ideas that have not been clad in a mathematical form. Whichever the reasons for the influence of the nominalistic concept, today it forms the basis for most algorithms and neural networks designed to model mental processes. Notwithstanding, in the area of the theory of intellect, the attitude of nominalism diverts the thought *from* seeking adequate mathematical concepts—algorithms and neural networks built on the nominalistic concept that disregard a priori content of the intellect come to impasse. And today, tracing the relationships between philosophical and mathematical theories of the intellect and outlining future research directions, we move away from Occam, who stands near the roots of scientific objectification, toward the idealistic realism of Plato and Aristotle, explaining the possibility of mind by combining apriority and adaptivity on a realistic basis. Understanding the relationships among the fundamental philosophical concepts of realism and nominalism is paramount for the following analysis of psychological and mathematical concepts.

3.3.2 Nominalism, Behaviorism, and Cybernetics

Psychological and philosophical analysis of mind, combining a philosophical tradition with an empirical aspect of the contemporary scientific method, sparkled at the beginning of the

twentieth century with the discoveries of Freud and Jung. However, combining philosophy and science met with difficulties. The success of the mathematical method (in the area of material substance) had advanced a requirement for objectification of scientific inquiry. In the empirical sciences, the only criterion of objectivity was seen in the reproducible experiments. Theoretical deductions and the very possibility of synthetic judgments a priori[8] were questioned. A priori concepts started losing ground, became lowered to the level of (at best) unproved hypotheses, and I would say that in some areas of science the temptation of objectivity eliminated the possibility for complicated synthetic a priori judgments—the beauty of the Platonic way of thinking was lost, and the possibility of scientific thinking, dressed not in the strict language of mathematical computations, seemed compromised.

Near the end of the nineteenth century, psychology started utilizing laboratory experimental procedures trying to approach the rigor of physical sciences. Many psychologists were impressed by the theory of reflexes rigorously advanced by Pavlov. However, any hope of utilizing objective scientific methods in the field of mental processes seemed impossible due to their complexity—a situation that was exacerbated by the discovery of the unconscious, which complicated the theory of mind. In this atmosphere, to resolve the dilemma between the objectivity and depth of investigation, behaviorism was born, a new scientific direction redefining psychology as a science of human behavior and an accompanying intellectual and philosophical movement.

A concept of behaviorism that attempted to explain the entire human psychology as a sequence of stimuli and reflexes and denied a need for consciousness in understanding of the intellect dominated American psychology from about 1920 to 1960. One of the reasons for the past popularity of behaviorism was a striving toward scientific strictness in the absence of mathematical methods adequate for the complicated problem of the analysis of mind. Seeing the only criteria of scientific objectivity in reproducible experimental result, behaviorism had to forgo considerations of deep mental processes. Behaviorism as a scientific school, as a temporary idealization of a complicated problem, created a scientific methodology of experimental psychology. It established the importance of the environment as a determining factor in human behavior, showed that the role of mental factors is often incorrectly exaggerated in everyday life, and successfully described multiple aspects of behavior in terms of external factors alone. However, behaviorism as a philosophy maintaining that the concepts of *consciousness*, *free will*, and *idea* are not needed in psychology and should be discarded, exerted an inhibiting influence on the development of concepts of mind.

Today, behaviorism is not popular; however, concepts of consciousness based on a sequence of inner stimuli and reflexes and concepts of learning based on environmental manipulation are still being formulated. To the extensive discussion of this topic I allow myself to add that a determinism by external factors, which is obviously important in human behavior, is not the most interesting aspect of it—what is most fascinating about human behavior is its spiritual aspect. Plato did not argue the fact that Socrates was killed in the material world. The great discovery of Plato was that there is a more important reality of human psyche, where Socrates exists today.

In the wealth of experimental data, collected by behaviorist scientists, cognitive scientists see confirmations of complex neural structures and of mental processes being important factors in human behavior, and conclude that the theory and philosophy of behaviorism are inadequate. However, notwithstanding the seeming arbitrariness and voluntary occurrence of the philosophy of behaviorism—as an attempt to reduce psychology exclusively to

external empirical factors—it is a continuation of the ancient philosophical tradition of nominalism expressed in psychological terms of the twentieth century.[9] And the schism between behaviorism and mentalism[10] is but one of many in the 2300-year debate between nominalists and realists.

Mathematical methods applicable to the field of psychology have been developed in the twentieth century and applied to modeling behavior and mind of living beings. These methods for the analysis of complex systems, such as factor analysis and stochastic process theory, describe much more complex phenomena than classical mathematical methods of mechanics and electrodynamics.[11] Nevertheless, physical systems inspiring the development of the new mathematics were very simple as compared with living beings. Factor analysis, developed by Spearman and Thurstone for the analysis of statistical correlations in multidimensional spaces, models stochastic or random deviations about the mean value. The mean value in factor analysis is defined by a single multidimensional deterministic phenomenon.[12] For example, a diversity of human abilities is modeled by many factors characterizing abilities. But the limitation of factor analysis by a single deterministic phenomenon is that a distribution of each ability among the population is characterized by random deviations from a single common mean value of a corresponding factor (that is, by a single mode).[13] Analogously, a theory of stochastic processes developed by Wiener models a single particle randomly moving under the bombardment of molecules in a drop of water—a very simple system compared with the mind or even with a mindless living being.[14]

Notwithstanding, the new mathematics was different in principle from calculus: it addressed the problem of multiple interacting bodies, and had an applicability far beyond its origin. Applying a new method to the problem of control in animals and machines, Wiener created a new science that he called cybernetics.

The emergence of cybernetics proceeded under the influence of the dominating psychological concept of behaviorism, which can be seen from the cybernetics' program paper (Rosenblueth et al., 1943). The authors defined teleological behavior as a purposeful behavior with feedback and emphasized that their understanding of teleology did not contradict deterministic behavior. Although noticing that the relationship between teleology, so defined, and a concept of freedom was problematic, they maintained that no qualitative differences had been found so far between animals and machines.

The influence of behaviorism in cybernetics has been and still remains strong, while originators of cybernetics have been ambivalent about the importance of mental processes and complex a priori internal structures. If Wiener's discussion of the relationships between his concepts and those of the empiricism of Locke and Hume is followed, it is seen that Wiener transcends the nominalistic basis of empiricism in relating universals to internal brain structures. Also, Rosenblueth et al. (1943) suggested that animals utilize predictive models of low orders, whereas humans, possibly, are capable of utilizing higher order predictive models. Later, Wiener came to appreciate the mathematical difficulties behind this suggestion, and he emphasized that using higher order predictive models is inadequate for the description of complex nonstationary systems, because of insufficient data for learning (Wiener, 1948).

Our approach to resolving this difficulty of insufficient data uses complex models with nontrivial a priori knowledge. The prior knowledge "makes up" for insufficient data and enables learning from a small amount of data. Wiener could come close to this concept, but he stopped short of incorporating nontrivial a priori knowledge into the mind's internal

models. This requires new, more powerful methods of nonlinear mathematics. Although the importance of nonlinear interactions for self-organization was clear to Wiener, who discussed this issue in *Cybernetics* (1948), nonlinear mathematics for self-organizing behavior did not yet exist. The nominalistic direction of thinking, it seems, distracted him from the systematic investigation of mathematical concepts of internal models. Is it related to the nominalism of empirical and behaviorist roots of cybernetics, or to the mathematical complexity of this step requiring new nonlinear mathematical methods?

Cybernetics achieved a combination of adaptivity with a priori knowledge based on internal adaptive models. By utilizing relatively simple models of stochastic stationary processes (Wiener filters), cybernetics combined adaptivity (of model parameters) with apriority (of models). In simple situations, the cybernetic method of system control based on a model led to optimal speed of adaptation, that is, to learning from a minimum amount of data. However, these models were too simple for modeling of mind. The simplicity of these models fitted the behaviorist concepts in psychology, thus reinforcing the two-way mutual influence of cybernetics and behaviorism and interfering with the development of complex internal models utilizing nontrivial a priori knowledge.

While discussing the inevitable shortcomings of early cybernetics, we should emphasize its most important achievement: the shift of emphasis from energetic and metabolic processes characterizing the material substance toward informational processes, which since remain the main subject of study for many researchers of the intellect. An informational theory of the intellect is a mathematical apparatus of the physics of mind, which is being created today. Among its challenges there is the 2300-year-old problem of combining adaptivity and apriority of mind.

3.4 Mathematics vs. Physics

3.4.1 Pythagoras, Descartes, Newton

Using mathematics to derive ultimate truths dates back to the Chaldeans and ancient Greeks. Ancient mathematicians were developing the rational thinking process among a predominantly mystical collective prerational consciousness that colored mathematical concepts. For pythagoreans, mathematics was not limited to formulating and proving theorems; they considered it the origin and cause of all beings. For example, they considered the number *one* to be also a force of unity, or the number *two* to be the cause of differentiation and decay. Prescientific, prerational attitudes ascribe mystical and spiritual properties to numbers—mathematics was full of mystic significance, connected to all mysteries of the universe. Pythagoreans, in Jungian words, projected onto mathematical objects contents of their unconsciousness.

The first person to formulate consciously the thought that mathematics should be used as a tool for elucidating philosophical truths in a rational way was Descartes (1637).[15] Whereas alchemists of the preceding centuries were seeking explanations of the properties of soul in matter, Descartes separated matter and spirit. By analyzing the phenomenological world, Descartes concluded that properties of mind cannot be rationally explained by properties of matter, and he postulated the existence of two types of substances, material and spiritual. By excluding all properties of matter except its spatial extent and reducing matter to a

mathematical expression, to movements governed by mathematical laws, Descartes freed matter from materialized residues of the idea of emanation—"little spirits flattering through the air." This simple geometric model of matter was a crucial step toward creating a condition in which man was able to foresee. He saw in mathematics an objective unification of a priori ideas and empirical data. The method of Descartes harmonized the relationship between the a priori and empirical on the basis of mathematics, in the definition of the subject of science as a search for a priori laws expressed in a mathematical apparatus combining a priori knowledge and empirical data.

Descartes did not succeed in his principle of finding in mathematics all philosophical truths. The scientific method of Descartes, in the area of material substance, was realized in the discoveries of Newton, who developed calculus, the mathematical apparatus adequate for combining the a priori and empirical. The Newtonian explanation for the motion of the Moon and the following successes of the computational (mathematical) approach in the eighteenth century established mathematical modeling as the ultimate scientific method. Since Newton, physical theory has been searching for computational (mathematical) models of nature,[16] and computational models became inseparable parts of the physical intuition of the world. A mathematical apparatus of physics created the possibility for predictions connecting a priori concepts (the first laws of physics) to empirical data in a way that is beyond doubt (that is, a priori). In philosophical terminology this is called synthetic a priori judgments, that is, complicated predictions that validity is of a priori origin (and not entirely a consequence of experimental measurement).

In the area of the spiritual substance, realization of the scientific method of Descartes, however, had not succeeded in the eighteenth century because the mathematical apparatus of calculus could not be used for modeling complex systems—even a three-body problem was too complex for exact analysis—much less forms of life or an organic function—the transcendental spirit. Attempts to create a unified theory of physics, biology, and psychology continued during the nineteenth century, when outstanding scientists conducted interdisciplinary research. Maxwell created the electromagnetic theory in physics and developed the trichromatic color theory of vision in psychology. Helmholtz discovered the law of energy conservation, a concept of free energy, and a theory of vortexes in liquids; in neurophysiology he measured the speed of electrical signals in axons; and in psychology he discovered mechanisms of adaptation in visual perception. These mechanisms turned out to be much more complicated than the laws of classical Newtonian physics. Helmholtz found that the perception of color at each point in a visual field is determined not only by the color spectrum at the point, but also by the averaged color of the entire visual field in a complicated nonlinear fashion. Color perception, in contemporary language, is described by a nonlinear and nonlocal field theory. Analyzing the process of visual perception as a whole, Helmholtz concluded that the perception process is unconsciously affected by previous perceptions. Thus, a nonlinear, nonlocal field of visual perception turns out also to be nonstationary. A mathematical apparatus of such field theory did not exist at the time. But for scientists of theoretical predisposition of mind a mathematical apparatus is a language of physical intuition and the only possibility of synthetic judgments a priori. Consequently, Helmholtz and other scientists concentrated their interests in those areas of physics in which existing mathematical methods could be utilized.

Psychological and philosophical analyses of mind continued without mathematical support until the creation of cybernetics and mathematical methods for analysis of complex

systems, including methods related to electronic computers. And only today do we see the beginning of the discovery of mathematics that combines the apriority and the adaptive empiricism of the spiritual substance.

3.4.2 Computation: Metaphor vs. Physical Model

> Time that is intolerant
> Of the brave and innocent,
> . . . .
> Worships language and forgives
> Everyone by whom it lives
> —AUDEN

A notion of computation is often used today for mental processes, sometimes in a direct sense and sometimes indirectly, as a metaphor. Metaphors play an important role in extending our knowledge, in the process of creativity, and it is possible that the most important discoveries are those in language. Analysis of word etymologies shows that language grows by metaphors and that this process leads to the growth of consciousness, for language is our most important tool of conscious understanding of the world.

A review of various metaphors that were used for mind can be found in Daugman (1988). One of the first to utilize a computational metaphor of the mind was Turing, who compared mind to a computer. Turing, in his famous test,[17] explored a controversial and provocative idea about the power of computers. Newton, on the contrary, did not suggest computational metaphors for physics and did not regard movements of celestial bodies to be similar to a computation. Newton was searching to create a computational method that would satisfy his intuition about how the world actually is. A principal difference between the two usages of computation is that Newtonian usage led to a successful description and prediction of the essential properties of movements of celestial bodies and other phenomena, but the Turing test did not predict properties of mind. Turing created a foundation for a theory of computation, and he challenged us to explain mind as a computation, to create a mathematical theory that could explain the mind. A mathematical apparatus for the physics of mind has to possess a certain elusive property called physical intuition. Whenever a mathematical model has been successful in the sense of Newtonian mechanics, it represents a significant aspect of the picture of the world as a specific physical intuition about Nature. According to our analysis in this chapter, the minimal ingredient of the physical intuition of mind is a combination of apriority and adaptivity. Although pure mathematical intuition seems to represent the aspect of mind that is related to a priori objects in the Platonian world of ideas, the physical intuition, together with a mathematical model that embodies it, possesses both aspects of mind; it combines the a priori with the empirical.

Physical intuition changes when new discoveries reveal unforeseen physics. There are many examples of a combined mathematical model and physical intuition replacing with time an old one: Gallilean relativity vs. Einstein's special theory of relativity, Newtonian mechanics vs. quantum mechanics, Newtonian gravitation vs. Einstein's general theory of relativity, thermodynamics vs. statistical physics, chemistry vs. molecular physics, quantum

electrodynamics vs. string theory, etc. A computational or mathematical model became an exemplary scientific paradigm, an ultimate, even if not final, physical understanding. According to this understanding, mathematical models of intellect aim to be not metaphors but the physical intuition of mind.

3.4.3 Physics of Mind vs. Physics of Brain

A mathematical apparatus of the physics of mind is being created today in parallel with the development of the physics of brain, or neurophysiology, which studies a material architecture and energetic processes forming the foundation of the mind. Temporary separation of the two fields is a necessary physical idealization, which is similar to Newton's creation of mechanics accounting not for quantum and relativistic properties of space. A peculiarity of the contemporary situation is our knowledge of the existence of the material structure of the brain, while our knowledge is insufficient for the creation of a unified theory.

Even if we knew the exact wiring diagram of all neurons in the brain, by itself this would not bring physical understanding of what is mind. The basic, fundamental principles of mind would still have to be found. Looking toward the future unified physical theory of mind and matter, physics of mind is being created today *to some extent* independently from physics of brain. The question of where to draw the line between the two areas is a subject of heated debate and disagreement among various schools studying the processes of cognition. The followers of strong AI[18] believe that the material and energetic structure of the brain play no principal role in the theory of intellect, but many neural network researchers believe that the theory of intellect can be developed only by modeling the neural structure of the brain. A choice of the right level of separation between physics of mind and physics of brain, the choice of the physical idealization, belongs to physical intuition and might turn out to be crucial for the creation of the physical theory of mind.

3.5 Kant: Pure Spirit and Psychology

Descartes founded the scientific method on the basis of rational thinking, the possibility of which was undoubtedly obvious to him. Rejecting Descartes' views, Kant discovered limits of rational thinking. By analyzing antinomies of reason, such as between causality and freedom, he concluded that certain truths are beyond rational understanding. Discovering antinomies of reason and limits of rational thinking, Kant was the first to use the rational method for a systematic study of the concrete content of a priori knowledge. Considering cognition as a meeting of a priori forms and sensory data, he continued the ancient Aristotelian tradition and specified a number of a priori forms and categories, such as space, time, unity, and multiplicity. In the critique of rational thinking, Kant reaffirmed the rational, scientific approach to elucidating the truths: origins of the antinomies of reason, he suggested, ought to be searched in the complicated nature of our a priori knowledge.

In his investigations of *the explicit content of a priori knowledge*, Kant systematically addressed the most perplexing problem of understanding of mind since antiquity, the problem of the contention between Aristotle and Plato at the philosophical level and between Chomskyan linguists at the abstract-computational level. Investigation of the explicit

content of a priori knowledge at the neurocomputational level, initiated by Grossberg, today draws interest from psychologists, neurobiologists, physicists, and mathematicians.

By overturning the understanding of the relationships between reason and content of mental processes, the philosophical criticism of Kant brought the abstract philosophical analysis of pure spirit close to the psychological philosophy coming 100 years later, which Nietzsche would write about: "Future philosophers . . . will become psychologists." He created preconditions for the nexus between the philosophy of pure spirit and the scientific method. However, Kantian understanding of interactions between reason and the outer world, including the metaphysics of natural science, contained numerous contradictions[19] that limited the influence of Kantian intuitions on the development of science. According to Kant, pure reason is the faculty of transcendental[20] or a priori cognition, and its contents are the a priori principles of mind. Kant separates the subject-Self that contains reason from the unknowable substance of the inner being, which he considers inaccessible to cognition and substantially indistinguishable from the unknowable outward object. By considering the a priori content of pure reason to be finite and amenable to complete knowledge, while assigning an infinite, forever impenetrable mystery to the nature of outward objects, to a thing-in-itself, Kant differs from contemporary psychologists and from contemporary physicists. Contemporary physics considers outward objects to be amenable to cognition and penetrable to the scientific method, and psychologists, following Dostoevsky, Nietzsche, and Jung, assign much more depth to the psyche, the foundation of which is in the collective unconscious, in the primary matter of Avicenna, and which does not seem to be penetrable to the scientific method (Chomsky, 1972).

Recall, for example, the way Kant begins his exposition of the concept of time by denying its empirical origin and affirming its a priori nature; but he then returns empirical reality to time and denies its transcendental reality. Despite contradictions of Kant's views and their opposition to contemporary ones, I think that his intuitions of the physics of mind not only should not be repudiated, but can and ought to be utilized in future mathematical concepts of intellect. A key to such understanding is provided by the Jungian theory of projection. Kant, in Jung's words, has projected unconscious depths of his psyche into the outward object, thing-in-itself, accomplishing by this projection, according to Jung's theory, the first step toward understanding. Completing the circle of understanding, today we internalize Kant's projection by understanding Kant's thing-in-itself as the primary matter of our psyche—the collective unconscious. In Jung's theory, considering the process of cognition as alternating projections and introjections, I discern the possibility of interpreting Kant's views on the nature of space and time. Namely, the a priori intuition of space and time exists at a definite depth of our psyche—in the region between consciousness and unconscious. Being an a priori category of the individual conscious mind, time is not transcendent with respect to the entire depth of our psyche; that is, in our psyche there are primordial layers of timeless perception that are not touched on by a concept of time, which does not penetrate the entire depth of the collective unconscious that consists on a larger part of timeless and spaceless archetypes.[21,22]

By internalizing the thing-in-itself, tying it up to the material substrate of the psyche, that is to the brain, the infinity of Kant's theory is extended in both directions, inside and outside, elucidating the penetrations of Kant's intuition: conscious principles of the mind are but a thin layer covering the impenetrable thing-in-itself of our unconscious. Who may know if future discoveries in physics would not uncover the infinite nature of the outward

object-in-itself? And the doubt is due to whether contradictions exist at the level of intuition and to where the infinity will occur—inside us or outside us.

3.6 Freud vs. Jung: Psychology of Philosophy

The old philosophical controversy between realism and nominalism, transcendence and immanence, adaptivity and apriority, was replayed again in the field of psychology during the age of great psychological discoveries at the end of the nineteenth and beginning of the twentieth century. By discovering the unconscious using inductive scientific method, Freud prepared the way for connecting the philosophical realistic concept of mind with the scientific world of matter. He identified modular structures within the psyche, such as *ego*, which includes conscious processes, and *id*, which includes unconscious processes. In an early unpublished work, Freud (1895) described psychic structures in terms of a system of interacting neurons. Three modular neuron systems were identified: the perceptual system ϕ, the memory system ψ, and the ω system, related to the consciousness. The properties of the synaptic connections in ϕ are genetically determined and contain a priori information, whereas adaptation occurs in ψ where synaptic properties can be altered by a transmitted signal. According to this theory, ego is described as a totality of the ψ cathexes or, in contemporary neural network terminology, by a set of the excited neurons in ψ with strong synaptic connections. This seems to be the first attempt at a neuronal explanation of the intentional states.

A number of concepts fundamental to Freud's theory of psyche have been inspired by the concept of neuronal organization of the brain. Later Freud uses a concept of cathexis to denote the investment of libidinal energy. While analyzing neuronal organization of brain, Freud conceives the concepts of the psychic energy and flight from the stimulus, the future Eros and Tanatos, the Life and Death instincts. In his published works, Freud did not refer to his earlier project of neuronal-based psychology: the significance of his psychological discoveries has surpassed an immature understanding of the organization of the brain and the physics of neural networks that has inspired his psychological theory. A similar exchange of ideas between physics of mind and physics of brain is also needed today, as our understanding is not sufficient for creating a unified theory of the two substances, matter and spirit.

An interplay of the two factors, apriority and adaptivity, that has been influencing philosophy for more than 2000 years, can be traced in Freud's views as well. In relating psyche to the brain, Freud asserts a reality of psychic processes. He identifies the generation of hallucinations as a primary psychical process, and processes involving inhibition by the ego he identifies as secondary. Unconscious processes in the brain, closer to the material basis of psyche, are given a more fundamental role, comparatively to the conscious processes in the ego. Thus on the one hand, Freud begins his analysis from the viewpoint of materialization of psyche, which seems to be close to philosophical realism, according to which mind has a real, a priori nature. But on the other hand, the role of the genetic a priori content of the psyche in his theory is relatively trivial and consists in differentiating between the afferent (sensory) and efferent (hallucinatory) signals. While discussing the importance of the match between the afferent signals coming from ϕ and the efferent

signals coming from ψ for the elementary cognition process, Freud does not mention any role of a priori content of the psyche in the shaping of the efferent signals. On the contrary, Freud emphasizes the role of individual memory, leaving mostly instincts to genetics and explaining mental processes on empirical grounds, as a suppressed experience, and not on an a priori basis, Freud moves away from the philosophy of realism toward nominalism (according to the philosophy of nominalism, concepts of mind are learned by generalizing individual experience). A salient aspect of Freud's theory is an attempt to account for the content of the psychic structures on a nominalistic basis—as a suppressed experience. In this, some contemporary critics see the root of the psychoanalytic diagnostic deficiency: the primordial genetic contents of the psyche, the collective unconscious ideas that are common to all of us and that are made conscious during psychoanalytic treatment, are misidentified as individual childhood memories.

A concept of unconscious discovered by Freud has been furthered by Jung, who discovered the existence and importance of primordial structures in our psyche, the collective unconscious. By analyzing types of problems and questions posed by thinkers of all ages and all peoples, from ancient Greeks to alchemists, from ancient Chinese to contemporary poets, he found something that unites the search for spiritual meaning. He called these somethings the archetypes of psyche's unconscious. Although the contents of archetypes are not accessible directly to the consciousness, they are manifested as it were by providing a framework for the consciousness. In the theory of archetypes I see an affinity of the concepts of psychology and of Aristotelian a priori Forms of the pure spirit—a priori forms-archetypes, meeting with individual experience, form concrete concepts of mind—in this way a possibility is created for a scientific approach to mind based on the philosophy of realism.

The philosophical schism between realism and nominalism, according to Jung (1934), to a significant degree has been due to the antagonism between two different psychological types: the introverted and extroverted. Introverted thinkers are more conscious about their internal thoughts and tend to emphasize a priori internal knowledge, whereas extroverted thinkers tend to emphasize learning from experience. Every "philosopher is secretly guided and forced into certain channels by his instincts" (Nietzsche). And to repeat once more, the scientific investigation is in the search for a synthesis of a priori knowledge with the adaptive learning from experience, the search started by Aristotle. The synthesis of apriority and adaptivity should transcend the limitations of attitudes of individual scientists, through the development of an individual psyche toward being equally conscious about both: the inside world of a priori concepts and the outside world of empirical events.

3.7 Wither We Go from Here?

3.7.1 Apriority and Adaptivity

The problem of combining adaptivity and apriority is fundamental to computational intelligence as well as to understanding human intelligence. There is an interrelationship among concepts of mind in mathematics, psychology, and philosophy that is much closer than currently thought among scientists and philosophers of today. From the contemporary point of view, the questions about mind posed by ancient philosophers are astonishingly scientific.

A central question to the work of Plato, Aristotle, Avicenna, Maimonides, Aquinas, Occam, and Kant was the question of the origins of universal concepts. Are we born with a priori knowledge of concepts or do we acquire this knowledge adaptively by learning from experience? This question was central to the work of ancient philosophers and medieval theologists, and it was equally important to theories of Freud, Jung, and Skinner. The different answers they gave to this question are very similar to the answers given by McCulloch, Minsky, Chomsky, and Grossberg.

During the 2000 years following Plato and Aristotle, the concept of apriority was tremendously strengthened by the development of a monotheistic religion in Europe, to the extent that it interfered with empirical studies. At the end of the scholastic era, human spirit felt strong enough to question a priori truths on empirical ground. Occam rejected the concept of apriority; he held nominalistic views that are opposite to realism. *Nominalism* considers ideas to be just names for classes of similar empirical facts. Occam prepared the way for the empiricism of Locke and Hume that is among the foundations of the scientific method. Jung has explained the schism between philosophies of realism and nominalism as due to two types of deep-seated psychological attitudes. In particular, nominalism and empiricism are related to an extroverted psychological attitude, which is at a premium in our pluralistic society. Thus it is not a coincidence or chance that nominalism continues to exert a significant influence on present scientific concepts.

The crisis in the field of early neural networks coincided with the contemporaneous downfall of behavioristic psychology and philosophy. Behaviorism, as a philosophy, impoverished the study of mind and was rejected in the 1960s. The downfall of behaviorism was but a milestone in the age-old debate between realism and nominalism. We saw that emergence of cybernetics proceeded under the influence of behaviorism. Similarly, behaviorism influenced early neural network research in the 1950s and 1960s. The concept of learning from examples without a priori knowledge did not follow the realistic philosophical direction outlined by McCulloch, but pursued the nominalistic philosophy together with behaviorism. And the downfall of early neural network research is related to its association with behaviorism and nominalism, a philosophy untenable any longer as a metaphysics of mind.

Tracing the metaphysical origins of our mathematical concepts of intellect is helpful for understanding not only the past of neural network research, but also the future. In particular, two concepts due to Aristotle have been examined. One is Aristotelian logic conceived to describe eternal truths. Another is the Aristotelian theory of mind describing adaptive, changeable Forms. The mathematical difficulties we are facing today can be traced to a contradiction in the Aristotelian treatment of these concepts. This contradiction is related to the Aristotelian disagreement with Plato, and to the Aristotelian rejection of Plato's Ideas for the new concept of Form. For 2000 years philosophers-realists, followers of Plato and Aristotle, analyzed *ontological* differences between Plato's Ideas and Aristotelian Forms, but the principled *epistemological* difference was not noticed. Ontology refers to existence: whereas Plato assumed that his Ideas exist in a separate world, Aristotle considered Forms as existing in our mind. Epistemology refers to the ways in which knowledge is acquired: in Plato's theory, Ideas are unchangeable eternal truths, whereas Aristotelian Forms are dynamic entities. Only recently, when Aristotelian logic was applied to mathematical modeling of mind, the contradiction between Aristotelian logic and theory of mind led to difficulties, contradictions, and an impasse. Analyzing the original contradiction will help us understanding future directions in the research of mind.

Aristotle developed a concept of Forms having, on the one hand, a universal a priori reality like Plato's Ideas, and on the other, being adaptive dynamic entities. Forms unite the worlds of spirit and matter. They exist a priori as potentialities in the world of spirit and they exist as actualities inseparably from individual objects in the world of matter. Adaptivity of mind is due to a meeting between the a priori Form and matter, forming an individual experience. This process explains learning. The major departure of Aristotelian theory from Plato's Ideas was that before a Form meets matter it exists as a potentiality; thus, it is not in its final form of a concept; it becomes a concept in the process of experience. This theory was further developed by Avicenna (XI AD), Maimonides (1190), Aquinas (XIII), and Kant (1781), among many other philosophers during the last 2300 years.

But Aristotelian logic is unsuitable for describing Forms, because Aristotelian logic deals explicitly with the eternal truths in their final crisp forms of concepts. *Crisp* is the opposite of fuzzy,[23] as defined by the law of contradiction or "excluded third," which is a central law of Aristotelian logic. According to the law of contradiction, every statement is either true or not true, and there is no third alternative. All statements and concepts of Aristotelian logic are crisp in that they obey this law. It might be appropriate for the eternally valid truths, but it is not applicable to our everyday intelligence, nor to fluid and adaptable Aristotelian Forms describing the process of learning. Since Aristotelian logic is the foundation of most of our algorithms, including rule-based AI, the difficulties and contradictions of rule-based systems are traced to Aristotle. We have discussed that the original contradiction of Aristotelian logic between the law of contradiction and the inherent uncertainty of human thought was historically resolved in two ways. Formal logicians from Boole to Russell rejected the uncertainty. The founder of fuzzy logic, Zadeh, rejected the law of contradiction. Fuzzy logic is needed for the Aristotelian theory of Form—theory of mind. Thus, the 2300-year-old contradiction between theory of mind and logic is resolved with fuzzy logic.

It is of interest to note that of the two concepts of McCulloch discussed previously, the one that became widely accepted by the neural network community was the concept of a neuron as essentially a binary device. This concept is obviously related to the crisp Aristotelian logic and its formalized contemporary descendant, Boolean calculus. Fascinated by the analogy between biological neurons and Boolean formalism, McCulloch and Pitts (1943) postulated simplified neurons having just two crisp states. McCulloch followed Aristotle in the two conceptual directions we are concerned with: logic and combining adaptivity with apriority, and thus inherited the original contradiction. It is instructive to note that when McCulloch countered the general trend, he foresaw the future direction of the science of intellect even though it was not pursued by his immediate followers. But when he followed the trend, his idea enjoyed immediate but relatively short-lived success. The Boolean property of early "formal" neurons was immediately accepted, but rejected since: it is now well understood that gradient learning as well as other learning methods in complicated nonlinear neural networks require smooth transfer functions. Artificial neural networks utilize smooth, fuzzy transfer functions. However, the nexus between fuzzy logic and Aristotelian concept of Forms, combining apriority and adaptivity, was not addressed.

3.7.2 Fuzzy Logic, Models, and Neural Fields

Let us summarize the results of our discussion so far. The fundamental issue in pattern recognition and computational intelligence was and continues to be the relative role of a

priori knowledge and adaptive learning. Computational intelligence techniques that utilize only one of these two factors face severe limitations. These limitations have been manifested in three different ways for the three types of algorithms. For algorithms based on the factor of adaptivity alone, the limitations have been manifested by the combinatorial training requirements. And for algorithms based on the factor of apriority alone, the limitations have been manifested by the combinatorial explosion of the complexity of rule systems. To overcome these limitations, model-based techniques have been developed for combining adaptivity and apriority, but they often lead to a combinatorial explosion of computational complexity.

The human intellect combines the two factors of apriority and adaptivity. According to the philosophical analysis dating to Aristotle, adaptive learning is based on a priori Forms. In today's mathematical language, adaptive parametric models come closest to Aristotelian Forms. A meeting between the a priori Form and matter can be understood mathematically as adaptive estimation of model parameters from the data. Thus, a successful approach to modeling human intellect ought to combine adaptivity and apriority in a model-based paradigm. However, algorithms that have been used in the past to combine adaptivity and apriority in a model-based paradigm lead to combinatorial computational complexity and are not suited for this purpose. The answer to the conundrum of combinatorial complexity requires understanding of this difficulty.

We saw that the analysis of the contradiction between Aristotelian logic and theory of Form provided us with the direction to look for a solution to this riddle. The major point of Aristotelian criticism of Plato's Ideas was that before a Form meets matter, it has to be *not in its final form of a concept.* But Aristotelian logic that underlies our algorithms, the Boolean calculus, and calculus of predicates that are based on Aristotelian logic operate with final crisp forms of concepts. The same is true about geometric models of model-based paradigms. It is the need to consider multiple combinations or associations between the concepts and the empirical world (signals, images) in the process of recognition that leads to combinatorial explosion. The answer to the puzzle of combinatorial complexity must be sought in overcoming Aristotelian logic that underlies our algorithms. Fuzzy logic founded by Zadeh in 1965 may provide keys to the answer. Intermediate computational steps (before Forms meet matter) should employ fuzzy representations of concepts. Neural networks with their inherent capability for fuzzy logic at the intermediate computational steps emerge as a vehicle for this new computational concept. Existing neural networks, however, lack the capability for representing complicated a priori knowledge. Current approaches to combining neural and logical processing do so by eclectic means of combining the old computational concepts in hybrid systems, while "new computational concepts are needed" (Sun and Bookman, 1995). Thus, researchers of intellect should be looking for a nexus of model-based and neural network concepts. How could this Aristotelian mathematics of mind be achieved? It seems that the future mathematics of mind will utilize complicated a priori Forms represented as fuzzy, spatiotemporal models in a neural architecture. A dynamic spatiotemporal model combining spatial information representation and temporal adaptive processing in a neural architecture is called a neural modeling field. The a priori neural modeling field is fuzzy. In the process of learning or adaptation it "meets" matter and becomes a less fuzzy or crisp concept. Such systems described in the second part of this book evolve by learning from external stimuli based on a priori models; it is an a priori adaptive fuzzy paradigm of Aristotelian mathematics of mind.

NOTES

1. Marvin Minsky has contributed to a number of approaches to computational intelligence. He is most famous for his contribution to rule-based AI or expert systems based on logical rules, which is referred to in this book's designation of the Plato–Minsky method.
2. It should be noted that notwithstanding the 2000-year dominating position of the Aristotelian physics based on eternal principles, Aristotle assumed the eternity of the world as an unresolved issue: "In areas where we have no proof" (*Topics*, 104b,15). Also, compare to Note 2 in Chapter 2.
3. Philosophers emphasized ontological differences between Aristotle and Plato, that is the nature of *being* of Aristotelian Forms vs. Plato's Ideas. Whereas Plato placed Ideas in a separately existing world of Ideas, Aristotle placed Forms into our minds. This ontological difference, much discussed throughout ages, is unimportant today because of the fundamental widening of our understanding of the diverse nature of *being*. Existence of genes, which persists through millions of years or existence of quantum fields and superpositions of quantum states are very different from existence of objects of everyday perception. And even our understanding of the latter was fundamentally changed by Kant. However, the difference in epistemology (the origin of knowledge, the learning) of Plato and Aristotle escaped philosophical scrutiny. For Plato, an explanation of learning was an afterthought (much as for the rule system developers). For Aristotle, the existence and origin of knowledge were equally important and had to be explained together.
4. The transcendence of God refers to His a priori nature, which transcends all experience. The immanence of God refers to His immediacy and relevancy at every moment in the ever-changing world, which in the contemporary language of the mathematical theory of intellect is called adaptivity.
5. Adaptive nature is a necessary aspect of the personal Deity.
6. Scholastic tradition has explored many concepts important to the understanding of the nature of intellect, which are too complicated for the analysis by available mathematical concepts. Notable among them is the concept of Trinity, mathematical analysis of which is beyond the scope of this book.
7. A scientist is the subject of scientific inquiry. Its object is an object of science. Objectification of the subject is equivalent to objectification of the object; both assume that subject and object exist independently from each other in an "objective" way.
8. Synthetic judgments a priori, according to Kant's terminology, are nontrivial (synthetic, non-tautological) conclusions derived from fundamental (a priori) truths in such a way that their validity is undoubted (that is, of a priori origin). According to Kant, mind possesses special abilities (of inborn, a priori origin) to form synthetic judgments a priori.
9. The psychological and philosophical position of Skinner, one of the leading behaviorist, is interesting to follow in light of Jung's theory of psychological roots of contradictions between realism and nominalism. Denying a nominalistic relationship, Skinner believes that a contradiction between nominalism and realism for Ockham has been purely linguistic and can be resolved by substituting *contingencies of behavior* for *properties* (Skinner, 1974, p. 94). In this denial of a principled (not linguistic) difference between nominalism and realism, one can trace a psychological attitude analyzed by Jung (see Section 3.6), confirmed by an affinity of concepts of nominalism and behaviorism as an extroverted attitude toward psychology. A philosophical affinity of behaviorism and nominalism can be traced in Skinner's discussions about "theories . . . supplement facts" (Skinner, 1974; p. x), or from Skinner's definition of a concept as a "set . . . which exist in the world" (ibid., p. 95), and from numerous other discussions in his book. His discussions of free will, mental processes, and many other concepts contain unproved conclusions and implicit unstated postulates, which is acceptable as a personal

position, but cannot be accepted as a scientific truth. This reveals that the basis of behaviorism as intellectual movement is psychological rather than scientific.

10. Mentalism maintains that complicated mental processes are essential for understanding human behavior and mind. Mentalism, which opposes behaviorism, is accepted by cognitive science.
11. Mathematical models of systems containing large numbers of particles have been created in classical physics as well, for example, the gaseous state equations, a computation of the speed of sound, and statistical physics in general. However, successes in these areas have been due to the fact that modeling of parameters of interest required considerations of statistical average states of a single particle, or sometimes correlated states of two particles. The problems of intellect are characterized by inadequacy of considering statistically averaged states and by a need to consider correlated states of all (or many) neurons in their interactions at every moment.
12. The mathematical foundation of Factor analysis is a linear technique based on a single Gaussian distribution (Thurstone, 1947). Gaussian distributions are adequate for the analysis of variability in physical systems with *a single* underlying deterministic phenomenology that is characterized by the mean values of the observables, while the variability is dominated by multiple random factors (Cramer, 1946). When the variability is due to *multiple* deterministic processes in addition to the random variability, Gaussian distributions are not applicable even approximately, and multimodal statistical distributions have to be utilized (Perlovsky, 1991a). Until recently, mathematical methods for using multimodal distributions were not available. Multimodal statistical distributions are considered in Chapter 4. Combinations of multimodal distributions and physical or dynamic models are considered throughout the book.
13. Thus, factor analysis is fundamentally limited in modeling diversity of human intellect.
14. The pointed similarity between Factor analysis and the Wiener processes comes from the underlying physical assumptions of a single deterministic process. The differences are in that factor analysis addresses a multidimensional structure of a statistical system, whereas Wiener's theory describes the system's dynamics.
15. There is a point of view that between Aristotle and Descartes the philosophical thought was dormant: "In the Middle Ages, as in childhood, opinions were formed subject to external authority . . . the motive for independent inquiry was weak" (Dixon, 1943). An opposite point of view (Jung, 1951) holds that great truths about human nature have been discovered by gnostics, theologists, alchemists, and philosophers during the time between Aristotle and Descartes, the key to which has been lost during the Enlightenment, at the beginning of the scientific age. Only now is our scientific thinking ready to attempt reclaiming this lost knowledge. [The return of scientific thoughts of today to philosophical concepts of the past, the breaking into today of the philosophical concepts of the preceding centuries, confirms Hegelian intuition about the nature of the development process (Hegel, 1807). An understanding of this process of spiral development in terms of physical–computational properties of mind would be an intriguing object of future analysis.]
16. An area of *computational* physics, which emerged recently, is involved in the development of more efficient computational methods within frameworks of the existent mathematical models. *Computational methods* in this book always have as meaning the mathematical principles–models of Nature. This differentiation is not an essential one—computational methods utilizing new principles and enabling solutions of qualitatively new types of problems lead to new physical understanding.
17. The Turing test is a thought experiment: a computer or a human is placed in a closed room and communications (questions and answers) are transmitted, say by a teletype. If as a result of such an interaction, it is not possible to determine if there is a human or a computer in the room, then mind is similar to a computer.
18. Strong AI, a term introduced by Searle (1980), designates a belief that the specific material structure of the brain is not essential for the understanding of mind.

19. Solovyov considered Kant's metaphysical foundations of the natural sciences as having "doubtful philosophical significance" (1885), and Nietzsche considered certain aspects of Kant's philosophy greatly confused: "Kant . . . —really lead astray" (1886; 5), "Kant's famous . . . error" (1887; III,6).
20. *Transcendent and transcendental* in this book corresponds to *transcendental* in the *Critique of Pure Reason* (Kant differentiates these notions).
21. The timeless and spaceless nature of archetypes of the collective unconscious is discussed by Jung (1951), who assumed the primordial origin of archetypes and their uniform nature among various peoples. A relatively recent origin of our a priori intuitions about ordered time and space is indicated by observations of psychologists (Lèvy-Bruhl, 1910) and linguists (Whorf, 1936) concerning different conceptions of time and space in different peoples; in particular, primitive consciousness perceives time and space as not quite ordered globally with a higher level of the local orderliness (in the local region of time and space where the tribe currently lives).
22. The epistemological problem of the relationship between the psyche and the outward object has not been solved. Who is right, Kant or physics? This question is far from being answered. It is impossible to accept any of the existing theories of the growth of scientific knowledge, or the received instant rationality of falsificationism, or Kuhn's irrationalism, or Lakatos' rationality of continuous growth (see, e.g., Lakatos and Musgrave, 1970), for none of these theories addresses the fundamental issue of the relationships between the growth of science and the content of a priori knowledge.
23. Crisp concepts obey the law of contradiction. Fuzzy logic deals with fuzzy concepts that do not obey the law of contradiction. An Aristotelian crisp concept is a limiting case of a fuzzy concept (in the limit of no fuzziness). Mathematical methods of describing adaptive fuzzy concepts are addressed throughout this book.

BIBLIOGRAPHICAL NOTES

This section lists references that were not identified in the text directly.

Einstein's quote is from "Remarks on Bertrand Russell's Theory of Knowledge" in *The Philosophy of Bertrand Russell*, P.A. Schillp, ed. Northwestern University Press, 1944, pp. 278–291.

Aristotle: Forms being a formative principle in an individual experience (*Metaphysics*); Forms-as-potentialities (*On the Soul*, III,4); the end cause and the formal cause (*Metaphysics*, Δ,24).

Avicenna on combining philosophy and theology (Avicenna, XI).

Maimonides on combining philosophy and theology (Maimonides, 1190), on finite angels (ibid., II:XII).

Aquinas on quidditive and existential aspects of mind (Aquinas, XIII).

Occam on universals, thoughts, and language (Occam, XIV).

History of philosophy (Windelband, 1883; Goodman, 1977).

A priori contents of mind: contents of pure reason (Kant, 1781); contents of the language faculty (Koster and May, 1981); contents of the visual system (Grossberg, 1976, 1980a,b, 1982, 1983, 1988, 1995).

Kant on the concepts of time and space [Kant, 1781, I(II)].

Nietzsche on unconscious guiding the consciousness (Nietzsche, 1886, 3).

Jung on the rational feeling function (Jung, 1921); on theory of projection (Jung, 1934); on archetypes of the collective unconscious (Jung, 1934).

Behaviorism: proponents (Watson, 1913; Skinner, 1974); critique (Jaynes, 1976; Grossberg, 1988b).

Factor analysis (Spearman, 1904; Thurstone, 1947).

Cybernetics (Rosenblueth et al., 1943; Wiener, 1948).
Wiener on empiricism of Locke and Hume (Wiener, 1948, Chapters 5 and 6); on mathematical difficulties of using higher order predictive models related to our analysis of combinatorial explosion (Wiener, 1948).
Scientific method and Descartes (Sutcliffe, 1968, in foreword to Descartes, 1637).
Learning with rule systems (Minsky, 1975, Section 6.1.1; Winston, 1984; Koster and May, 1981; Botha, 1991; Bonnisone et al., 1991; Keshavan et al., 1993); in Chomsky's linguistics (Botha, 1991).
Neural a priori structures (McCulloch and Pitts, 1943)
Newton's biography (More, 1934).
Grossberg on afferent and efferent neural signals (Grossberg, 1980a).
Mathematical analysis of difficulties of algorithms and neural networks built on the nominalistic concept is given in Chapter 2; also see Perlovsky (1994a; 1998a).
On language, metaphors, and consciousness (Jaynes, 1976; Daugman, 1988).
Fuzzy logic (Zadeh, 1962).

PROBLEMS

The following problems list topics for oral presentations and papers. Each topic can be addressed at varying levels of detail and scope of effort. A one-week study should concentrate on one or two reference sources documenting the proposed points. A semester-level effort may include a literature survey and result in a conference presentation or a journal paper. Also, a group of students can be assigned a topic to prepare a group discussion.

3.1–1 Survey Plato's concept of Ideas as described in Parmenides. (Note that the words Ideas and Forms are used interchangeably by translators.) List examples and properties emphasizing the a priori and universal properties of Ideas. Document adaptive properties of Ideas, if any. Analyze similarly Minsky's concept of models (Minsky, 1968b: *Matter, Mind, and Models*). Compare Plato's and Minsky's views. Prepare a journal paper.

3.2–1 Survey Aristotelian concepts of Forms as described in *Metaphysics*. List separately examples and properties emphasizing the a priori nature of Forms (Forms-as-potentialities, formal causes) and those emphasizing the adaptive, dynamic aspects and the role of Forms in learning (Forms-as-actualities, end causes). Prepare a journal paper. Extend this study toward other Aristotelian works, including *On the Soul*.

3.2–2 Survey the Aristotelian concept of the *end cause* as described in *Metaphysics*. Compare it with the concept of intentionality of mind and intentional states as described by Brentano and Searle. (See Searle, 1980 for further references.) Prepare a journal paper.

3.2–3 Review Aristotelian books on logic: *Analytics* and *Topics*. Document a priori, eternal, and unchangeable properties of those concepts, to which the laws of logic are applicable. Document Aristotelian opinions on this matter. Seek evidence for the opposite, adaptive properties of Aristotelian logic, if any. Prepare a journal paper.

3.2–4 Review Maimonides' concept of finite angels (1190: *The Guide for the Perplexed*). Document a need for this concept as a connection between the a priori spirit and matter. Compare finite angels to neural signals coming from the brain to the muscles and to the sensory organs (e.g., eyes). Compare this with Grossberg's concept of efferent signals

(1980a: How does a brain build a cognitive code?). Compare this with your favorite artificial intelligence paradigm if any. Prepare a journal paper.

3.2–5 Review Aquinas' concepts of sensation and intellection (XIII: *Summa contra Gentiles*). Concentrate on his description of an interaction between the a priori, universal and material, individual. Compare this with Grossberg's concept of efferent and afferent signals (1980a: How does a brain build a cognitive code?). Compare this with your favorite artificial intelligence paradigm if any. Prepare a journal paper.

3.3–1 Review Occam's concept of terms (XIV: *Summa logicae*). Concentrate on his description of the relationships between the universal and individual. Compare this with (an opposing) Chomsky's concept of language faculty (Botha, 1991: *Challenging Chomsky*). Prepare a journal paper.

3.3–2 Review the influence of behavioristic concepts on the development of early cybernetics (Rosenblueth et al., 1943; Wiener, 1948; Skinner, 1974). Prepare a journal paper.

3.4–1 Document Kantian views on the a priori knowledge as described in the Introduction to the *Critique of Pure Reason* (1781). Compare this with Minsky's concept of a priori knowledge (1968b: *Matter, Mind, and Models*). Prepare a journal paper.

This page intentionally left blank

part two

MODELING FIELD THEORY

New Mathematical Theory of Intelligence with Engineering Applications

This is the main part of the book. It consists of seven chapters describing modeling field theory (MFT) and its engineering applications. Chapter 4 presents the theory of modeling fields and describes neural networks based on this theory. Chapter 5 describes the maximum likelihood neural network (MLANS) and its applications to grouping and recognition. Chapter 6 describes the Einsteinian neural network and its applications to signal and image processing. Chapter 7 describes applications to prediction, association, tracking, and sensor fusion. Chapter 8 discusses the possibility that quantum computations are performed by biological neurons and describes quantum MFT. Chapter 9 addresses fundamental mathematical limitations on learning. Chapter 10 establishes relationships between MFT and Kantian theory of mind, and describes an organization of Kant–MFT intelligent systems.

This page intentionally left blank

MODELING FIELD THEORY

The challenges of modeling the mind discussed in the previous chapters can be summarized as follows. The mind is capable of combining complicated a priori knowledge with adaptive learning in the presence of perpetual uncertainties of a diverse nature. Algorithms and neural networks designed to model intelligence face a combinatorial computational explosion, indicating that they are not suitable for this purpose. The two most important concepts of modeling the mind, complicated a priori internal representations and learning from experience, have been developed independently of each other. This was emphasized by Minsky: "theories of learning and theories of representation were developed independently." And the roots of this divide have been traced to Aristotle. In his theory of mind, Aristotle rejected the ready-made eternal Platonian Ideas. However, Aristotelian logic was created for ready-made eternally valid concepts, not for fluid, fuzzy Forms. Fuzzy logic promises a resolution of this impasse, if a way to combine it with apriority and adaptivity can be found.

A theory of neural modeling field developed in this chapter accomplishes just that. It combines complex, structured a priori knowledge of an internal model with adaptivity of the model parameters. And it avoids combinatorial explosion by using fuzzy logic. In other words, modeling field theory is a long sought for representation of an a priori model, which supports both learning and adaptation.

A concept of similarity between the internal model and the world is central to the mathematical theory of mind based on an internal model. We consider first a general approach to constructing similarity measures that are suitable for complex models of the world (composed of multiple local models). Three types of similarity measures are considered: Aristotelian, Zadeh's (fuzzy), and adaptive fuzzy. Dynamic equations of modeling field theory are obtained that maximize the adaptive-fuzzy similarity between the model and the world. Two instantiations of this general theory are developed using the fundamental principles of likelihood and information: a Bayesian similarity measure leading to maximum likelihood learning and Shannon's similarity measure leading to maximum information or maximum entropy learning.

Modeling field theory provides a mathematical apparatus of fuzzy adaptive logic for Aristotelian Forms, which are represented as dynamic neural fields. It is an intelligent system with an internal instinct or drive to increase the similarity between the internal model and the world. That is an instinct to learn.

4.1 Internal Models, Uncertainties, and Similarities

4.1.1 Certainty and Uncertainties

Any order that we perceive in the world is incomplete, partial, and interspersed with uncertainty. It even seems sometimes that any order in the world is imposed by us: early concepts of beauty in human societies are associated with order and symmetry; we are often fascinated with order emerging naturally, such as in crystals; and even most abstract works of art seems to impress us with an order revealed in chaos. Might it be that the appeal of order is related to our desire and ability to foresee? At the same time, complete order without variations is often perceived as stale and lifeless. Our ability to perceive both order and variability seems to be essential. Thus, both ought to be represented in our mind.

An order is a natural property of mathematical models. Simple mathematical functions whose shapes are determined by few parameters are an epitome of order. More complicated mathematical constructs are used to represent uncertainty and disorder. A well established mathematical theory of uncertainty is the theory of probability. Its empirical roots are related to a specific type of uncertainty, the uncertainty of chance, the uncertainty of events whose relative frequencies can be predicted. The theory of probability deals with *random* variables, whose random properties are essential and inherent to the modeled phenomena. Other types of uncertainty include unknown, but nonrandom deterministic quantities. Fuzzy variables and fuzzy logic is a recently emerging mathematical technique for modeling uncertainties that cannot be characterized by their frequencies of occurrence. (There are also chaotic processes, which are unpredictable even though they are characterized by known and nonrandom quantities; their unpredictability is due to exacerbated inaccuracies inherent in predictions. In this book we will not be much concerned with those.)

4.1.2 Models and Levels

The functioning of mind requires three fundamental abilities: an ability to use concepts of perception and cognition, which are represented by internal models; an ability to establish a correspondence between the models and the world; and an ability to generate behavior. A mathematical description of the first two abilities is the main topic of this section. It seems natural that behavior is generated after a correspondence is established between the internal representation and the world. It turns out, however, that there is a specific type of behavior, which is inseparable from the other two abilities, the behavior of adaptive learning. So, even an elementary "unit of intelligence" has to combine all three abilities.

Establishing a correspondence between the internal representations and the world is a nontrivial mathematical issue. For example, consider the following rule: if CHAIR then SIT. To implement such a rule, a correspondence must be established between CHAIR and a subset of the sensory data about the world. But, a logical concept CHAIR is incommensurate with the world. At best you can find a word "chair" in a book page, but you are not supposed to sit on a book page! It follows that model-concepts should have multiple levels of representation, including levels corresponding to written signs and to the objects in the world.

Modeling field theory developed in this chapter describes mathematically neural interactions among three levels: a higher level, in which an individual chair in the world is

represented internally by an output activation signal from the "concept-CHAIR"; a lower level, in which the bodily sensory expectations are generated including an expected visual representation of an image-CHAIR; and an a priori level, in which CHAIR is represented as a parametric model-CHAIR. The "concept-CHAIR" activation signal is an input data to other models including higher level models. There, in the interaction with other models, this signal is interpreted by the rest of the mind as indicating the presence of a chair in the scene (for example, a model SIT will direct our body to sit on this object). The lower level image-CHAIR is "matched" to the sensory visual input images. Sensory signals and expected images are "in the same domain"; they are commensurate, and a similarity measure can be defined between two commensurate quantities. The a priori parametric model-CHAIR is a part of the mechanism used in recognition and adaptation, which is responsible for generating the image-CHAIR. It is interesting to compare the above description to the intelligent tracker in Section 1.1.4 (Problem 4.1-1).

4.1.3 Lower Level Models

Let us consider internal models that represent "world" in the sensory domain. Could these models be just pictures or movies of the entire scenes? No, such models would hardly be useful for recognizing objects of interest within the scene. Instead, the models have to represent objects of interest. An object of interest could be a significant part of the entire scene (say, a room), or a small part of the scene (say, a chair). This leads to a need for multiresolutional models. Let us now concentrate on a single resolution level. To be able to recognize multiple objects (such as chairs, tables), the model has to be compositional, it has to contain multiple models of objects. Internal models often used in the model-based vision are detailed physical models of objects of interest that require no adaptation or learning. They are suitable in a highly ordered environment without variabilities. When unpredictable variabilities are expected, adaptive models have to be used. Adaptive models are parametric functions, with unknown parameters that represent nonrandom deterministic uncertainty. We say that a deterministic parametric model $\mathbf{M}(\mathbf{S})$ with parameters $\mathbf{S}$ models data $\mathbf{x}$, if

$$\mathbf{x} = \mathbf{M}(\mathbf{S}) \tag{4.1-1}$$

Boldface is used to emphasize that the corresponding quantities are vectors (or arrays, or sets of numbers). If the data $\mathbf{x}$ are the output of sensory cells, $\mathbf{M}$ are perception-concept models; if the data $\mathbf{x}$ are excitations of other models farther along the processing levels, $\mathbf{M}$ are more complicated object models, or cognitive-concept models. To distinguish a number of different objects or concepts, index k is used: $\mathbf{M}_k(\mathbf{S}_k)$, $k = 1, \ldots, K$. Pieces of data (input nodes) are numbered by index n, $\mathbf{x}_n$. For a black–white image, the data are pixel intensities; for a color image, the data $\mathbf{x}_n$ are a vector or set of intensities in several color bands; at a higher processing level, $\mathbf{x}_n$ could be vectors of intensities or features extracted at lower levels. In neural terminology, $\mathbf{x}_n$ are activation degrees of a subset of the input nodes (e.g., axon rate of firing of retina cells). Model parameters, $\mathbf{S}_k$, may describe position, orientation, and pose of the object. For speech recognition, $\mathbf{x}_n$ could be individual signal samples or a set of signal samples from a small window, or features of phonemes, and $\mathbf{S}_k$ describes articulation of a vocal tract, etc. Or, $\mathbf{x}_n$ could be a set of features extracted from the data in

the result of feature extraction or preprocessing: for example, edge detection is performed by an early visual system, and could be performed by an artificial retina. Complex systems, in general, have multiple levels, and concepts or classes identified at each level become input data at the next level.

According to the three levels of models, we use the term model in three ways: the concept k-model is the activation output signal representing this concept to the rest of the brain; the a priori $\mathbf{M}_k(\mathbf{S}_k)$ model is a function of parameters, a process that generates the image-model-k; and the same notation, $\mathbf{M}_k(\mathbf{S}_k)$, is used for the image-model, which is *similar* to the data for specific values of $\mathbf{S}_k$. The data, $\mathbf{x}_n$, evoked by an object or concept n are activation degrees of a (large) number of input nodes at a given level, a field of activations. Correspondingly, the image-model is a simulated (predicted) value of this field and, in general, a model predicts different values for different pixels n, $\mathbf{M}_k(\mathbf{S}_k, n)$. If input data $\mathbf{x}_n$ are due to (evoked by) an object of class k, there are values of parameters $\mathbf{S}_k$ such that the data match the model

$$\mathbf{x}_n = \mathbf{M}_k(\mathbf{S}_k, n) \tag{4.1-2}$$

The image-model, thus, is a prediction of the data vector. When constructing models for robots, it is necessary to account for the properties of objects, sensory systems, and preprocessing (preprocessing is accomplished at the lower levels, between the sensor nodes and the considered model level). In reality, a perfect match as in Eq. (4.1-2) cannot be attained because there are multiple sources of uncertainty causing deviations between the model and the data.

Thus, a measure of similarity has to be introduced.

4.1.4 Similarity Measures

A similarity measure has to accomplish more than just measuring a degree to which Eq. (4.1-2) holds. It has to measure a similarity between the entire set of data (which is a continuous stream of overlapping patterns) and a set of models. Thus, a partition (association, segmentation) of the entire data set and a set of models has to be a part of the similarity measure. Such a measure is fundamental to model-based recognition based on internal representations. Before considering explicit functional forms of similarity measures in the next section, we first go over general approaches to defining such a measure. It has to account for a distributed representation of an object by a subset in the field of input nodes (synapses, pixels, or samples). Input data are denoted $\mathbf{x}_n$, $n = 1, \ldots, N$. Association (or segmentation) of the field of input nodes with objects can be described mathematically as a partition Ξ of pixels $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ into subsets ξ_k corresponding to particular objects, $\Xi = \{\xi_1, \ldots, \xi_K\}$. Subsets ξ_k are characterized by membership functions, $\xi(k|n) = 1$ if pixel n belongs to an object k, $n \in k$, or $\xi(k|n) = 0$ if $n \notin k$.

Given a partition Ξ and a set of model parameter values $\mathbf{S}_k$ for each object's model, a conditional similarity measure between pixel n and model k can be introduced, $\mathrm{l}\left[\mathbf{x}_n|\Xi, \mathbf{M}_k(\mathbf{S}_k, n)\right]$, or for simplicity of notations $\mathrm{l}(n|k)$. The role of partition here is to define the pixels belonging to object k, $\mathrm{l}(n|k) = 0$ for $n \notin k$; and, for $n \in k$, $\mathrm{l}(n|k)$ could be defined, e.g., as a function of $\left[\mathbf{x}_n - \mathbf{M}_k(\mathbf{S}_k, n)\right]^2$. On a par with $\mathrm{l}(n|k)$ we will consider its logarithm, $\mathrm{ll}(n|k) = \ln \mathrm{l}(n|k)$ (and we will refer to either of these as a similarity measure). The k-model

similarity, that is the similarity for all pixels belonging to k-subset (kth model) for a given segmentation, can be defined as a product (or sum) of conditional pixel similarities

$$\mathrm{l}(k|\Xi) = \prod_{n\in k} \mathrm{l}(n|k) \quad \text{or} \quad \mathrm{ll}(k|\Xi) = \ln \mathrm{l}(k|\Xi) = \sum_{n\in k} \mathrm{ll}(n|k) \tag{4.1-3}$$

The reason for using simple product or sum above is in that we are looking for the simplest expression satisfying an intuitive notion of similarity. This definition assumes that $\mathrm{l}(n|k) \geq 0$ [so that $\mathrm{ll}(n|k)$ is a real number]. Then, $\mathrm{l}(k|\Xi)$ is a positive function, which is close to zero when a similarity is low, and $\mathrm{ll}(k|\Xi)$ takes both positive and negative values, and it takes large negative values for a low similarity. Given the above definition, model parameters, conditional on segmentation, can be obtained by maximizing the conditional similarity

$$\max_{\mathbf{S}_k} \mathrm{ll}(k|\Xi) \Rightarrow \mathbf{S}_{k,\Xi} \tag{4.1-4}$$

A total conditional log-similarity for all pixels and all models for a given segmentation can be defined as a sum of model log-similarities

$$LL(\Xi) = \ln L(\Xi) = \sum_{k} \quad \mathrm{ll}(k|\Xi) \tag{4.1-5}$$

And the total unconditional measure of similarity LL is obtained by maximizing the above over all segmentations; this leads to the best segmentation and the best model parameter values:

$$LL = \max_{\{\Xi\}} \max_{\mathbf{S}_k} LL(\Xi) \Rightarrow \{\Xi, \mathbf{S}_k\} \tag{4.1-6}$$

Here, LL is a total measure of similarity between the input data set (a field of input nodes) and a set of models. And $\{\mathbf{S}_k\}$ is a set of parameters corresponding to the best match. When it does not introduce confusion, we call any L, LL, l, ll, max L, or max LL *a similarity measure*.

Aristotelian Similarity. The above measure of similarity is defined by using crisp partitions Ξ of the input data. Within each crisp partition, a definite set of pixels corresponds to each object with no uncertainty. Partitions, thus, obey the Aristotelian logic law of contradiction (or excluded third): each pixel either belongs to a representation of a particular object, or does not (there is no third alternative). Near the end of the nineteenth century, Aristotelian logic was mathematically formalized by effort of a series of outstanding mathematicians, including Cantor, Frege, and Russell. This work culminated in the famous Gödel theorems (discussed later in Chapter 11). The formalization of Aristotelian logic was based on identification of the logic and set theory: logical predicates are identified with subsets. A partition Ξ is equivalent to a set of predicates or statements of Aristotelian logic of the sort: pixel-n $\in$ object-k. Accordingly, the measure of similarity defined above we call an Aristotelian similarity, or A-similarity, A-L, or A-LL. Summarizing Eqs. (4.1-3) through (4.1-6), the A-similarity is defined as

$$\text{A-}LL = \max_{\{\Xi\}} \sum_{k} \quad \max_{\mathbf{S}_k} \sum_{n\in k} \quad \mathrm{ll}(n|k) \tag{4.1-7}$$

Finding parameter values that maximize A-similarity requires two steps: first each k-model similarity is maximized over the parameters of this model and second, the best parameters are selected that correspond to the best partition (segmentation). The first step can be approached by solving the equation

$$\partial/\partial \mathbf{S}_k \; \mathrm{ll}(k|\Xi) = 0, \quad \Rightarrow \mathbf{S}_{k,\Xi} \tag{4.1-8}$$

The second step requires consideration of a combinatorial number of partitions. Thus, the A-similarity has a high inherent combinatorial computational complexity, which is determined by multiple partitions (associations/segmentations), Ξ. Since every pixel could be associated with every object, the total number of partitions is on the order of $O\left(K^N\right)$, and the total computational complexity of A-similarity, C_{A} is on the order of

$$C_{\mathrm{A}} \sim O\left(K^N\right) \cdot C_{\max} \tag{4.1-9}$$

where $C_{\max}$ is a computational complexity of finding conditional maxima.

Fuzzy Similarity. Combinatorial complexity associated with segmentation can be eliminated by using fuzzy logic. In honor of Zadeh, the founder of fuzzy logic, we call it Z-similarity, Z-*L*. Instead of a crisp partition found by combinatorial search Ξ, Z-similarity is based on fuzzy class memberships $f(k|n)$, which may take any value within [0,1] determining a degree of pixel n belonging to an object k, and which are specified a priori (even if only approximately). How can we transform A-similarity into Z-similarity? Let us rewrite A-similarity in terms of crisp membership functions $\xi(k|n)$ as follows: rewrite $\Sigma_{n\in k} \; \mathrm{ll}(n|k)$ as $\Sigma_n \; \xi(k|n)\mathrm{ll}(n|k)$. If we substitute crisp membership functions $\xi(k|n)$ for fuzzy membership functions, $f(k|n)$, for the k-model similarity measure, in place of Eqs. (4.1-3) and (4.1-7) we obtain

$$\mathrm{ll}(k) = \sum_n \; f(k|n)\mathrm{ll}(n|k) \tag{4.1-10}$$

$$\text{Z-}LL = \max_{\{\Xi\}} \sum_k \quad \max_{\mathbf{S}_k} \sum_n \; f(k|n)\mathrm{ll}(n|k) \tag{4.1-11}$$

The parameters of each model are obtained by maximizing this model similarity,

$$\begin{aligned} \max_{\mathbf{S}_k} \text{Z-}LL \quad \text{or} \quad & \partial/\partial \mathbf{S}_k \sum_n \; f(k|n)\mathrm{ll}(n|k) \\ & = \sum_n f(k|n) \; \partial \mathrm{ll}(n|k)/\partial \mathbf{S}_k = 0, \quad \Rightarrow \mathbf{S}_k \end{aligned} \tag{4.1-12}$$

The Z-similarity has a low inherent computational complexity because of no need to consider multiple segmentations. The computational complexity of Z-similarity, C_{Z} is on the order of

$$C_{\mathrm{Z}} \sim O(N \cdot K) \cdot C_{\max} \tag{4.1-13}$$

A disadvantage of Z-similarity is due to nonadaptive fuzzy class memberships, which could lead to unduly fuzzy or unduly crisp concepts with class memberships that do not represent

actual uncertainty and unsuitable for representations of actual objects. This property makes Z-similarity too restrictive for learning or adaptive systems. (This limitation of fuzzy logic was emphasized by its founder Zadeh in 1997.)

Adaptive Fuzzy Similarity. We would like to define adaptive fuzzy similarity, or AZ-similarity, in such a way that it combines advantages of A- and Z-similarities: adaptive segmentation and low computational complexity. So, let us avoid using any predetermined segmentation, and define a similarity measure so that associations of data and models emerge in the process of parameter estimation that maximizes the similarity measure. Thus, originally every pixel has a chance to be associated with each model. Various models are possible alternatives for each pixel, therefore, pixel similarity $\mathrm{l}(n)$ is defined as a sum over alternatives, partial similarities $\mathrm{l}(n|k)$,

$$\mathrm{l}(n) = \sum_k \mathrm{l}(n|k) \quad \text{or} \quad \mathrm{ll}(n) = \ln\left[\sum_k \mathrm{l}(n|k)\right] \tag{4.1-14}$$

Note that here we sum similarities, not their logarithms. Of course, before specific functional expressions are specified, the difference between $\mathrm{l}(n|k)$ and $\mathrm{ll}(n|k)$ is just in the notations. Notations above are selected in such a way that if uncertainties are probabilistic, similarities can be interpreted as probabilities (or pdf): according to the rule of combining probabilities of alternative events, the probabilities should be added. According to this definition, the pixel similarity is large even if just one of the models predicts this pixel well. The total similarity is defined as

$$\begin{aligned} \text{AZ-}L &= \max_{\{\mathbf{S}_k\}} \prod_n \mathrm{l}(n) = \max_{\{\mathbf{S}_k\}} \prod_n \left\{\sum_k \mathrm{l}(n|k)\right\} \text{ or} \\ \text{AZ-}LL &= \max_{\{\mathbf{S}_k\}} \sum_n \mathrm{ll}(n) = \max_{\{\mathbf{S}_k\}} \sum_n \left\{\ln\left[\sum_k \mathrm{l}(n|k)\right]\right\} \end{aligned} \tag{4.1-15}$$

Here, parameters of all models $\{\mathbf{S}_k\}$ have to be estimated simultaneously: values of parameter for each model affect parameter values for other models. This accords with our desire for emerging segmentation. To see the emergence of segmentation, let us examine parameter estimation equations,

$$\partial/\partial\mathbf{S}_k \sum_n \mathrm{ll}(n) = \sum_n \partial/\partial\mathbf{S}_k \{\ln \mathrm{l}(n)\} = \sum_n [1/\mathrm{l}(n)]\, \partial/\partial\mathbf{S}_k \left[\sum_{k'} \mathrm{l}(n|k')\right] = 0 \tag{4.1-16}$$

Among $\mathrm{l}(n|k')$ only one item for $k' = k$ depends on $\mathbf{S}_k$: $\partial/\partial\mathbf{S}_k \left[\sum_{k'} \mathrm{l}(n|k')\right] = \partial/\partial\mathbf{S}_k\, \mathrm{l}(n|k)$. We rewrite this derivative by using an identity

$$\partial y = y \cdot \partial \ln y \tag{4.1-17}$$

Using this, we obtain,

$$\sum_n [1/\mathrm{l}(n)]\ \partial/\partial\mathbf{S}_k\, \mathrm{l}(n|k) = \sum_n\ [\mathrm{l}(n|k)/\mathrm{l}(n)]\ \partial \mathrm{ll}(n|k)/\partial\mathbf{S}_k = 0 \tag{4.1-18}$$

Here, the term $l(n|k)/l(n)$ approaches 1 if model k predicts pixel n much better than any other model, and it approaches zero, if model k predicts pixel n much worse than some other model. This term thus has properties suitable for the fuzzy membership function, and therefore we define

$$f(k|n, \mathbf{S}_k) = l(n|k)/l(n) = l(n|k)/[l(n|1) + \cdots + l(n|K)] \tag{4.1-19}$$

This expression is similar to the Bayesian a posteriori probability (1.2-15), and if uncertainties are probabilistic, and pdf are used as similarities, it can be interpreted as such. In this definition we explicitly emphasized the dependence of the fuzzy membership function on model parameters. These adaptive fuzzy memberships $f(k|n, \mathbf{S}_k)$ are not given a priori, but are computed from partial similarities, $l(n|k)$. The denominator here normalizes fuzzy memberships to be between 0 and 1 (partial similarities are positive) and implements a competition between concepts k for the datum n. The fuzzy memberships can be thought of as producing partial activations of the concept k (full activation of model k depends on partial activations of k from all input nodes n). Using this definition, estimation equations can be written as

$$\sum_n f(k|n, \mathbf{S}_k)\ \partial ll(n|k)/\partial \mathbf{S}_k = 0 \tag{4.1-20}$$

This is similar to nonadaptive similarity [Eq. (4.1-12)], however, here fuzzy class memberships are not defined a priori, they are functions of model parameters and the fuzzy segmentation emerges in the process of model estimation.

In the definition of AZ-similarity [Eq. (4.1-15)], the sum over k includes all alternative hypotheses or models k, for each pixel n. If we expand parentheses $\{\cdot\}$ in the AZ-L definition, it can be written as a sum of K^N items, each item being a conditional likelihood of the type of Eq. (4.1-5) for A-L defined for a particular segmentation Ξ. Thus, Eq. (4.1-15) contains all segmentations as alternatives. Still, AZ-similarity is noncombinatorial, as we see from parameter estimation [Eq. (4.1-20)]. The AZ-similarity measure combines advantages of A- and Z-similarities: adaptivity and low computational complexity. Similar to the complexity of Z-similarity [Eq. (4.1-13)],

$$C_{\mathrm{AZ}} \sim O(N \cdot K) \cdot C_{\max} \tag{4.1-21}$$

The factor $C_{\max}$ is somewhat larger in this case because fuzzy class memberships have to be computed.

Equation (4.1-20) leads to a mathematical description of the Aristotelian theory of mind: it combines apriority and adaptivity. In the next section we define a dynamic procedure for computing parameters $\mathbf{S}_k$ maximizing AZ-similarity, while concurrently leading to the emergent segmentation $f(k|n, \mathbf{S}_k)$ determined in the process of learning. A dynamic system implementing this procedure is called Modeling Field Theory (MFT).

4.2 Modeling Field Theory Dynamics

Modeling Field Theory, or MFT, is a system of dynamic equations maximizing AZ-similarity. Dynamics of MFT implements the Aristotelian theory of learning in the

following sense. According to the Aristotelian theory of Forms, learning is based on a priori adaptive Forms that are given a priori as potentialities and evolve into specific concepts in the process of learning. In MFT, learning starts from fuzzy Forms, given by highly fuzzy class memberships $f(k|n, \mathbf{S}_k)$ corresponding to a priori models with uncertain values of parameters $\mathbf{S}_k$, and learning proceeds toward deterministic concepts, given by low-fuzzy class memberships $f(k|n, \mathbf{S}_k)$ and final models with certain values of parameters $\mathbf{S}_k$.

4.2.1 Overview of the MFT System

Let us summarize the architectural organization of the modeling field theory (MFT). In a conceptual way, at some general level of approximation, it follows that of the brain. Brain combines modular heterarchical structure with multilevel hierarchical organization. Although multiple processing levels in the brain are well established, brain is not a uniform hierarchy where each level sends its results up to the next one in a final form. Instead, a significant interaction among many levels takes place during a process of concept formation and recognition. There is evidence that signal processing in the brain includes iterative loops encompassing several levels: e.g., word recognition affects phoneme recognition, etc. Thus, levels of processing can at best be identified only tentatively, and the brain is not a strict hierarchy, but a heterarchy. We call this heterohierarchical organization.

A single level of MFT is described in detail in this section. Most of the notations has been already introduced. At every level, input nodes (or pixels) n contain data $\mathbf{x}_n$; these are the degrees of activation of certain concepts at a lower level. At the considered level, these are input data, a field of activations. Models $\mathbf{M}_k(\mathbf{S}_k, n)$ predict (or model, simulate) these data, to the best extent possible, by modifying their parameters, $\mathbf{S}_k$. These adaptive models are emergent concepts. While adapting to input data, they compete with each other for evidence in the data. These models-concepts can be viewed as intelligent agents that often operate with a significant degree of independence.

At every level, a large number of emerging agents–concepts–objects is usually present and compete for evidence at any moment. So the entire input data set $\{\mathbf{x}_n\}$ is associated with a set of agents (objects or models) in a fuzzy way. A pixel (activation node datum) $\mathbf{x}_n$ may provide for a partial activation of the agent k, this partial activation is given by the fuzzy class membership $f(k|n, \mathbf{S}_k)$, or $f(k|n)$ for shortness. It is interpreted as a neural weight connecting input node n with agent k. And it is determined by a similarity between the datum $\mathbf{x}_n$ and its model-k prediction, $l(n|k)$. The model $\mathbf{M}_k(\mathbf{S}_k, n)$ predicts a large number (a modeling field) of individual data pieces $\mathbf{x}_n$. A similarity between $\mathbf{x}_n$ and $\mathbf{M}_k(\mathbf{S}_k, n)$ is denoted $l(n|k)$, which we call partial or pixel similarity. We freely use abbreviated notations whenever possible without causing ambiguities, e.g., arguments are omitted and indexes (n, k, etc.) are used in place of the corresponding quantities ($\mathbf{x}_n$, $\mathbf{S}_k$, etc.).

A mathematical apparatus of MFT relates various quantities introduced so far, and gives the learning/adaptation laws for the model parameters resulting in the maximal similarity. We will complete the description of MFT in several steps. First, AZ-similarity measure AZ-*LL* and the partial activations $f(k|n)$ have been already defined in terms of partial similarities $l(n|k)$, [Eqs. (4.1-15) and (4.1-19)]. Second, in the next section we formulate dynamic MFT equations maximizing AZ-*LL* in terms of $l(n|k)$ and their derivatives with respect to the models $\mathbf{M}_k$. Third, $l(n|k)$ will be explicitly specified as functions of the models, $\mathbf{M}_k$. And fourth, models will be explicitly specified as functions of parameters $\mathbf{S}_k$. The third

step is considered in Sections 4.3 and 4.4. We define two types of similarity measure, one based on the concept of probability or likelihood, leading to a Maximum Likelihood neural network (Section 4.3 and Chapter 5), and another, based on the concepts of information and entropy, leading to an Einsteinian neural network (Section 4.4 and Chapters 6 and 7). The fourth step, model specification, is application specific; a number of models will be introduced throughout the book for various applications.

4.2.2 MFT Dynamic Equations

Now, we derive the MFT system dynamic learning equations for the model parameters and partial activations (neural weights). These equations are derived by maximizing the AZ-similarity AZ-*LL*, using a variation of the gradient ascent method. According to the gradient ascent, at every moment t, parameters are changed incrementally, $\sim dt$, along the direction of the gradient of AZ-*LL*,

$$d\mathbf{S}_k = dt \cdot (\partial \text{ AZ-}LL/\partial \mathbf{S}_k) \tag{4.2-1}$$

Using Eqs. (4.1-15) through (4.1-19),

$$d\mathbf{S}_k/dt = \partial \text{ AZ-}LL/\partial \mathbf{S}_k = \sum_n f(k|n)\ \partial \text{ll}(n|k)/\partial \mathbf{S}_k \tag{4.2-2}$$

Dependence of $\text{ll}(n|k)$ on parameters $\mathbf{S}_k$ is only through the models $\mathbf{M}_k$; evaluating the derivative by the chain rule,

$$d\mathbf{S}_k/dt = \sum_n f(k|n)\ \partial \text{ll}(n|k)/\partial \mathbf{S}_k = \sum_n f(k|n)\,[\partial \text{ll}(n|k)/\partial \mathbf{M}_k]\,[\partial \mathbf{M}_k/\partial \mathbf{S}_k] \tag{4.2-3}$$

$$f(k|n) = \text{l}(n|k)/\left[\sum_{k'} \text{l}(n|k')\right] \tag{4.2-4}$$

Here model parameter updates are given as a weighted sum over the field of input nodes n. Considering this as a neuronal equation, $f(k|n)$ are interpreted as neural weights responsible for partial activation of model k due to input node n. Equations (4.2-3) and (4.2-4) define a convergent dynamic system of MFT, which always converges in finite time to a maximum of AZ-similarity, under the assumptions that AZ-*LL* is differentiable and finite. We consider the convergence proof in Section 4.6.

Conceptually, it might be more satisfactory to have the neural weights $f(k|n)$ defined by a dynamic equation, rather than as above. Such a dynamic equation can be obtained by taking a derivative of the above definition (see Problem 4.2-1)

$$df(k|n)/dt = f(k|n) \sum_{k'} \left[\delta_{kk'} - f(k'|n)\right] \text{ll}'(n|k')\ \mathbf{M}'_{k'} \cdot d\mathbf{S}_{k'}/dt \tag{4.2-5}$$

Here,

$$\begin{aligned} &\text{ll}'(n|k') = \partial \text{ ll}(n|k')/\partial \mathbf{M}_{k'}, \quad \mathbf{M}'_{k'} = \partial \mathbf{M}_{k'}/\partial \mathbf{S}'_k, \\ &\delta_{kk'} \text{ is 1 if } k = k', 0 \text{ otherwise} \end{aligned} \tag{4.2-6}$$

This equation also results in a convergent system, maximizing similarity AZ-*LL*. A convergence proof is considered in Section 4.6. These equations can be implemented by a single neuron for every k, or by a group of neurons, a subnetwork for each k, if models are complex.

4.2.3 Continuous MFT System

Modeling the mind with a concept of neural networks composed of individual neurons with isolated nodes is an approximation, which may be valid to varying degrees for various brain subsystems. There are many mechanisms in the brain that are more continuous than digital. First, a single synapse is not a mathematical point, but a continuous structure. Then, a neuron has a complicated tubular structure, within which complex processes are possible, and some researchers hypothesize that quantum-level computational processes might be involved. And, of course, there are essentially continuous processes involving changes in concentrations of certain neurally important chemicals on a scale much larger than a single neuron. Another motivation for considering continuous MFT is the possibility of engineering implementations other than digital computer. While quantum MFT is considered in Chapter 8, here we develop a continuous formulation.

First, let us consider a continuous field of input activations, so instead of n we will use a continuous index ν, while the index k referring to objects, concepts, and models, will be considered as discrete, $k = 1, \ldots, K$. Also, the number of parameters $\mathbf{S}_k$ for each model is still considered finite. Later we discuss possible roles for continuous fields of parameters and models.

Reformulation of the MFT equations for the continuous field of input data is straightforward. Equation (4.2-5) is not affected and in place of Eq. (4.2-3), we have

$$d\mathbf{S}_k/dt = \int d\nu \; f(k|\nu)\,[\partial \ln[l(\nu|k)]/\partial \mathbf{M}_k]\,[\partial \mathbf{M}_k/\partial \mathbf{S}_k] \tag{4.2-7}$$

Equations (4.2-5) and (4.2-7) define MFT over the continuous ν-field that converges to a maximum of the continuous modification of AZ-similarity:

$$\text{AZ-}LL = \max_{\mathbf{S}_k} \left\{ \int d\nu \; \sum_k l(\nu|k) \right\} \tag{4.2-8}$$

In certain situations, it is advantageous to consider continuous sets of models that depend on continuous sets of parameters. Deformable models of continuous geometric patterns such as murphs or snakes and appropriate models and measures of similarity can be defined so that the infinite number of models and parameters can be learned from finite training data.[1]

4.2.4 Heterarchy, Multiple Scales, and Local Maxima

In a multilevel MFT system, agent-concepts or classes emerging or activated at each level become input data at the next level. Multilevel, multiscale organization addresses computational complexity: resolution of each higher level in space–time decreases and scale increases. Roughly speaking, this results in an exponential savings of processing and memory requirements. If the resolution of each next higher scale decreases by s and scale increases by the same amount, then l levels would cover s^l larger scale in

space–time compared to the bottom level, and it would require only l times the processing and memory requirements compared to the bottom level. This type of multilevel and multiscale organization corresponds to our intuition about mind: we remember and can attend to a lot of details in the immediate surrounding, whereas on larger space–time scales we remember and attend to fewer details. It is also well known that certain events are retained in the memory with great detail for long time—this is evidence for heterarchy. Multilevel organization concerns not only memory but the entire conceptual organization of our mind: certain concepts (internal models) are more general and include a number of subconcepts (submodels or lower level models). But the same concept (model) can serve as a lower level or a higher level concept in various situations—again, evidence for heterarchy. A multilevel system requires a multilevel internal model, composed of a number of submodels operating at various levels. And mechanisms are needed for heterarchical partitioning-associating sensory data with various submodels.

Multilevel organization also helps in addressing a problem of local vs. global maxima, which plagues complicated nonlinear systems that operate on a single level. MFT uses two mechanisms to help find the global maximum. First is adaptive fuzziness of the models. High initial fuzziness helps avoid local maxima, by "not seeing" them. Second, in a multilevel system the problem of local maxima is properly resolved at higher levels. Local maxima of similarity correspond to misidentified concepts-objects, including missing concepts and false concepts. Wrong concepts are identified as such at a higher level, and corrective actions are taken, including resetting or redirecting activities at a lower level. Specific examples of such systems will be discussed throughout several subsequent chapters.

Let us formulate a mechanism of the agent-concept activation that is necessary for providing bottom-up input to the next level. The degree of the agent activation used as input at the next level ought to be determined by a degree of similarity between the agent-model and data. In other words, an agent activation is the degree of recognition of an object or concept. A simple mechanism of agent k activation level $a(k)$ is by summing partial activations $f(k|n)$ over the input data field,

$$a(k) = \sum_n f(k|n) \tag{4.2-9}$$

4.2.5 MFT, Fuzzy Logic, and Aristotelian Forms

Sections 4.1 and 4.2 described the general concepts of MFT, a learning system based on adaptive internal models of the world. MFT is a dynamic system, and its dynamic evolution constitutes the learning process. MFT learning is determined by two factors: the a priori internal models and real-time input data. Therefore, learning is combined with the a priori knowledge contained in the models. (Here "a priori" refers to all models that have been acquired before the current experience.)

MFT resolves the problem of the combinatorial complexity explosion faced in the past: MFT dynamics contains no combinatorics. The key to defining the noncombinatorial dynamics is the AZ-similarity measure, which combines a priori models with fuzzy logic and adaptivity. Fuzziness of MFT agent-classes or concepts is determined by the specific shape of partial similarity measures $l(n|k)$. They should be defined in such a way that at the beginning of the learning process, when parameter values are uncertain, the fuzziness

is high. And the fuzziness is reduced during learning, so that when parameter values attain their accurate values, the classes become crisp or low-fuzzy concepts. The matching of fuzziness to the uncertainty is important for efficient learning, avoiding local maxima. Let us repeat that this process implements the Aristotelian conception of learning: uncertain a priori Forms-as-potentialities become certain concepts in the process of learning.

Specific AZ-similarity measures are defined in Sections 4.3 and 4.4. They lead to specific types of neural networks. A general neural network architecture based on MFT dynamics is considered in Section 4.5. We call it MFT or Model-Based Neural Network. The AZ-similarity measures defined in the following sections lead to two types of neural networks: Maximum Likelihood (ML) Adaptive Neural System (ANS), or MLANS, and Shannon–Einsteinian or Maximum Entropy ANS (MEANS). A complete specification of MFT requires model definition (in addition to a similarity measure). Several types of models are considered in Chapters 5, 6, 7, and 9.

4.3 Bayesian MFT

A general theory of MFT developed in the previous section requires an explicit specification of a partial similarity measure, $l(n|k)$ in terms of models and data, and specification of models $\mathbf{M}_k$ ($\mathbf{S}_k$) in terms of parameters. In this section we develop a partial similarity measure using a fundamental concept of statistics, the Bayesian likelihood, and subsequent chapters will consider specific models. This similarity measure combines probabilistic aspects of similarity together with deterministic and fuzzy ones. Specific expressions for Bayesian A-similarity and AZ-similarity are obtained along with learning equations, maximizing the similarities. We call the MFT system of learning equations based on Bayesian AZ-similarity the Maximum Likelihood Adaptive Neural System (MLANS).

The maximum likelihood, or ML, is a fundamental statistical principle. In addition to the intuitively appealing notion of "the most likely," it has certain fundamental mathematical advantages. If the model accurately models the expected data, so that the deviations are random, the ML estimation is "the best" or close to the best one in the following mathematical sense: *the ML estimation is asymptotically unbiased and efficient*. The meaning of these words is as follows. *Asymptotically* means that the amount of data (training labeled data, or otherwise) is sufficiently large; sufficiently large means the limit of infinity, but often, this limit is closely approached (or exactly attained) for just a few data points per every model parameter. *Unbiased* means that on average, estimated parameters attain their true values. *Efficient* means that the average errors (standard deviations) of estimated parameters attain the lowest possible values. These lowest possible errors, as functions of the available data, can be estimated by using the Cramer–Rao lower bound, which is discussed in Chapter 9. Thus, the ML estimation leads (asymptotically, and often practically as well) to the fastest possible learning.

4.3.1 Bayesian A-Similarity Measure and the Principle of Maximum Likelihood

When all deterministic variabilities in the data are accounted for in the model $\mathbf{M}_k$ ($\mathbf{S}_k, n$), the deviations of data $\mathbf{X}_n$ from the model can be treated statistically. For each class, the

model should be developed so that it gives a statistical expectation (average value) of the data conditional on the class-hypothesis H_k,

$$\mathbf{M}_k(\mathbf{S}_k, n) = E\{\mathbf{x}_n|H_k\} = \int \mathbf{x}_n \text{pdf}(\mathbf{x}_n|H_k)\ d\mathbf{x}_n \tag{4.3-1}$$

Here, H_k includes class k, segmentation Ξ, model prediction $\mathbf{M}_k(\mathbf{S}_k, n)$, and model parameters $\mathbf{S}_k$, $H_k = [\Xi, \mathbf{M}_k(\mathbf{S}_k, n)]$, and the expectation $E\{\cdot\}$ is taken with respect to the H_k-conditional probability density function, $\text{pdf}(\mathbf{x}_n|H_k)$.

A pdf is called likelihood when considered at a given data value $\mathbf{x}_n$ as a function of model parameters. A concept of likelihood, as a measure of similarity between a model and data, is due to Bayes, therefore we will refer to similarity measures developed in this section as Bayesian similarities. We define a conditional partial similarity measure $\text{l}(n|k)$ as a joint likelihood of the data $\mathbf{x}_n$ and model H_k; the joint likelihood, $\text{pdf}(\mathbf{x}_n, H_k)$, accounts for the conditional likelihood of the data $\mathbf{x}_n$ given that we are observing class k, $\text{pdf}(\mathbf{x}_n|H_k)$, and for the likelihood of observing class k, $P(k)$:

$$\text{l}(n|k) = \text{pdf}(\mathbf{x}_n, H_k) = P(k)\text{pdf}(\mathbf{x}_n|H_k) \tag{4.3-2}$$

The second equality here is obtained by using the law of conditional probability, $\text{pdf}(a, b) = P(a)\text{pdf}(b|a)$, where a and b stand for the hypothesis and data, respectively. The probability of the hypothesis $P(k)$ is called in statistics a prior probability because in classical statistics it is considered known prior to data $\mathbf{x}_n$ observation. In our approach, $P(k)$ is a part of the model and can be adaptively learned similarly to other model parameters. Usually we model $P(k)$ with just one parameter, that is, we consider $P(k)$ a constant for each class or model. Because $P(k)$ is a proportion of observations (pixels) from class k, we call it class rate.

Combining the above equation with A-L definition [Eq. (4.1-7)], we obtain the Bayesian A-similarity

$$\text{A-}L = \max_{\{\Xi\}} \prod_k \text{l}(k|\Xi) = \max_{\{\Xi\}} \prod_k \max_{\mathbf{S}_k} \prod_{n\in k} P(k)\ \text{pdf}(\mathbf{x}_n|H_k) \tag{4.3-3}$$

Let us discuss when the total A-L similarity and Ξ-conditional similarities for classes $\text{l}(k|\Xi)$ above can be interpreted as the corresponding likelihoods. In particular, under what conditions the joint pdf of all pixels belonging to the concept k, $\text{pdf}(\{\mathbf{x}_n, n \in k\}, H_k|\Xi)$, factors into the product as above, $\prod_{n\in k} P(k)\ \text{pdf}(\mathbf{X}_n|H_k)$ and the joint pdf of all data and models conditional on the segmentation also factor into $\prod_k \text{l}(k|\Xi)$. First, this factoring of the joint pdfs into products occurs when deterministic models $\mathbf{M}_k(\mathbf{S}_k, n)$ accurately account for all sources of deterministic variabilities and remaining variabilities are random and uncorrelated among pixels. Thus, while modeling deterministic relationships among pixels, we assume their statistical independence. For uncorrelated data, joint pdfs factor into products of individual pdfs. Second, this factoring of the joint pdfs into products can be understood as follows: let parameters $\mathbf{S}_k$ include other pixels $\mathbf{x}_{n'}$ that may be correlated with $\mathbf{x}_n$; and let the model be given by the linear regression of $\mathbf{x}_n$ on $\mathbf{x}_{n'}$ and the deviations $(\mathbf{x}_n - \mathbf{M}_k)$ are uncorrelated with $\mathbf{x}_{n'}$ (see Problem 4.3-1). In this second case, we do not assume statistical independence among pixels. Therefore, the pdf of all the data $\{\mathbf{x}_n\}$ and models $\{H_k\}$ for a given segmentation (Ξ-conditional) can always be written as a product

over data n by designing proper models. The only assumption that we need to make is that the prior probabilities of models (class rates) are uncorrelated, that is

$$P(k = 1, \ldots k = K) = \prod_k P(k) \tag{4.3-4}$$

Then, the Ξ-conditional likelihoods (pdfs) are given by the above expressions:

$$\mathrm{l}(n|k, \Xi) = \mathrm{pdf}\left(\{\mathbf{X}_n\}, \{H_k\}|\Xi\right) = \prod_k \prod_{n \in k} P(k)\mathrm{pdf}\left(\mathbf{x}_n|H_k\right) \tag{4.3-5}$$

so that the A-L similarity is the likelihood, and the parameter estimation procedure defined by the Bayesian A-L similarity maximizes the likelihood. The best parameters for each object are determined by maximizing the Ξ-conditional similarities for classes, $\mathrm{l}(k|\Xi)$ over the parameters, and then selecting parameters corresponding to the best segmentation Ξ. Estimation of parameters by maximizing the likelihood is called the maximum likelihood (ML) parameter estimation.

The deterministic models $\mathbf{M}_k(\mathbf{S}_k, n)$ that accurately account for all sources of deterministic variability often lead to Gaussian class-conditional densities. If the deviations between the model and data are caused by multiple random effects, then, according to the Central Limit theorem,[2] the pdf often can be approximated by a Gaussian density:

$$\begin{aligned} \mathrm{pdf}\left(\mathbf{X}_n|H_k\right) &= G\left[\mathbf{X}_n|\mathbf{M}_k(\mathbf{S}_k, n), \mathbf{C}_k(\mathbf{S}_k)\right] \\ G\left(\mathbf{x}_n|\mathbf{M}_k, \mathbf{C}_k\right) &= (2\pi)^{-d/2}(\det\ \mathbf{C}_k)^{-1/2}\exp\left(-0.5\ \mathbf{D}_{nk}^{\mathrm{T}}\mathbf{C}_k^{-1}\mathbf{D}_{nk}\right) \\ \mathbf{D}_{nk} &= \mathbf{x}_n - \mathbf{M}_k \end{aligned} \tag{4.3-6}$$

where G is a Gaussian density with the mean given by the deterministic model $\mathbf{M}_k$ and the covariance $\mathbf{C}_k$, which is either directly estimated from the data or modeled. In certain cases, variabilities in the data are caused by specific physical mechanisms that do not satisfy conditions of the Central Limit theorem and Gaussian densities might not be appropriate; in these cases appropriate densities should be used. One general way to model deviations between the data and models, which is suitable for any statistical density of the deviations, is by utilizing mixture densities as considered in the next section. In this section, in addition to the general MLANS formulation, we also consider Gaussian densities for class-conditioned pdfs.

Let us summarize the various sources of variabilities and their modeling in MLANS. The models $\mathbf{M}_k(\mathbf{S}_k, n)$ describe deterministic relationships among data $\mathbf{x}_n$. Statistical dependencies, or correlations, are described by the covariance matrixes $\mathbf{C}_k$. Random probabilistic variabilities are represented by the pdf. Unknown deterministic variabilities are represented by the adaptivity of parameters $\mathbf{S}_k$. Fuzzy uncertainty about values of these parameters also can be represented using the covariance matrices. A simple way of accomplishing this is to specify large covariance matrixes at the beginning of learning process and let them be gradually reduced in the process of learning. More sophisticated techniques will be discussed throughout the book. (Note that strictly speaking, the enlarged covariances and resulting pdf are not statistical quantities as defined in Chapter 1; instead, they should be considered as fuzzy class memberships parameterized by $\mathbf{C}_k$.)

The above formulation of the Aristotelian similarity based on likelihood is a fairly broad approach to model-based pattern recognition. It addresses the top level of the problem, while it omits details important for specific application areas and for specific approaches to controlling the combinatorial explosion inherent in A-similarity. Most of techniques that have been discussed in the literature (such as multiple hypothesis testing algorithms) can be formulated within this framework. The need to consider all or many of the partitions or segmentations, Ξ, is the main source of the combinatorial complexity of the model-based pattern recognition based on A-similarity.

4.3.2 Bayesian AZ-Similarity Measure

Here we develop a Maximum Likelihood Adaptive Neural System (MLANS), an MFT paradigm using Bayesian likelihood for AZ-similarity measure. MLANS, like general MFT, considers a joint problem of concurrent segmentation, model estimation, and similarity maximization and solves this problem without combinatorial complexity. The joint likelihood is obtained by considering segmentation as a part of the model, which has to be estimated from data. So, segmentation emerges in the process of model estimation. A joint likelihood is built from individual pixel likelihoods, similar to the general MFT method: unlike what was found in the previous section, here the individual pixel likelihood, $l(n)$, is not conditioned on a segmentation and considers various classes as probabilistic alternatives. Therefore, the likelihood for each pixel is a sum of pdf of alternative model hypotheses:

$$l(n) = \text{pdf}(\mathbf{x}_n) = \sum_k \text{pdf}(\mathbf{x}_n, H_k) = \sum_k P(k)\text{pdf}(\mathbf{x}_n|H_k) \tag{4.3-7}$$

Hypotheses H_k here include the object-class k, models $\mathbf{M}_k$, and parameters $\mathbf{S}_k$, but do not include segmentation, which emerges in the process of model estimation. In some applications it is convenient to consider each object as a class (of pixels), and in other applications, multiple objects form a class; the following chapters consider both cases. The densities, $\text{pdf}(\mathbf{x}_n)$ and $\text{pdf}(\mathbf{x}_n, H_k)$, are the likelihood counterparts of the general concepts of partial pixel similarity $l(n|k)$ and total pixel similarity $l(n)$ considered in Section 4.1.3. Class rates, $P(k)$, are the probabilities, or expected relative frequencies of pixels "belonging" to various objects. Usually, they are not known a priori and should be estimated together with other parameters of the model. Class rates satisfy the constraint (see Problem 4.3-2):

$$\sum_k P(k) = 1 \tag{4.3-8}$$

The meaning of this constraint is that some object(s) among $k = 1, \ldots, K$ will definitely be encountered (with probability 1), or in other words, a signal in every pixel definitely originates from some of these objects. It follows that the set of K object models has to encompass all possible objects. In many practical cases this requires that we include among K a class of "other" or "unknown" objects with the appropriate uncertainty of their properties. In the following chapters we will discuss ways to accomplish this.

As discussed in the previous section, the total likelihood can be written as a product of pixel likelihoods. This is a general property of the model-based likelihood with properly selected models, and it does not assume independence among pixels. Let us repeat this

argument again, because of its importance: the factoring is due to relationships among pixels being included into the models [that is, pdf($\mathbf{x}_n|H_k$) may depend on $\mathbf{x}_{n+1}$ etc., which are included into a set of $\mathbf{S}_k$, if needed]. Thus, total likelihood is given by

$$L = \text{pdf}\{\mathbf{x}_1, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_N\} = \prod_n \text{pdf}(\mathbf{x}_n) = \prod_n \left\{ \sum_k P(k)\ \text{pdf}(\mathbf{x}_n|H_k) \right\} \tag{4.3-9}$$

The Bayesian AZ-similarity is defined as a maximum of the total likelihood,

$$\text{AZ-}L = \max_{\mathbf{S}_k} \left\{ \prod_n \sum_k P(k)\ \text{pdf}(\mathbf{x}_n|H_k) \right\} \quad \Rightarrow \{\mathbf{S}_k\} \tag{4.3-10}$$

so that the model parameters are obtained in the process of ML estimation.

As in the case of general AZ-similarity, sum over k in the above expression includes all alternative hypotheses or models $k = 1, \ldots, K$. If we expand the parentheses $\{\cdot\}$ in this expression, it can be written as a sum of K^N items, each item being a conditional likelihood of the type of $l(k|\Xi)$ in Eq. (4.3-3) defined for a particular segmentation Ξ. Thus, Eq. (4.3-9) contains all segmentations as alternatives. Probabilistic segmentation given by partial similarities emerges from Eq. (4.3-10) in the process of parameter estimation as discussed in the next section.

4.3.3 MLANS Learning Equations

Maximization of the likelihood [Eq. (4.3-10)] is achieved in MLANS in a similar way to the general MFT. One difference here is that we need to account for the normalization constraint for class rates (4.3-9). This is accomplished by the Lagrange multiplier method: instead of the likelihood, (4.3-10), or its logarithm, AZ-*LL*, we have to maximize[3]

$$\text{AZ-}LL' = \max_{\mathbf{S}_k} \left\{ \sum_n \ln \left[\sum_k P(k)\ \text{pdf}(\mathbf{x}_n|H_k) \right] + \lambda \left(\sum_k P(k) - 1 \right) \right\} \tag{4.3-11}$$

and to find λ to satisfy the constraint. If the class rates are known a priori, the additional item above is identically zero and does not affect estimation of other parameters. Using the gradient ascent method, similar to Eqs. (4.2-1) through (4.2-4), we obtain

$$d\mathbf{S}_k/dt = \sum_n f(k|n)\ \left[\partial \ln \left[\text{pdf}(\mathbf{x}_n|H_k)\right]/\partial \mathbf{M}_k\right] [\partial \mathbf{M}_k/\partial \mathbf{S}_k] \tag{4.3-12}$$

Here, the fuzzy membership of datum n in concept k or, alternatively, partial activations of concept-agent k by datum n, $f(k|n)$, are determined by partial similarities, as in Section 4.1,

$$f(k|n) = P(k)\text{pdf}(\mathbf{x}_n|H_k)\ /\text{pdf}(\mathbf{x}_n) \tag{4.3-13}$$

If the rates are not known, they are included into a set of parameters $\mathbf{S}_k$, and their estimation equation is derived similarly to the above (see Problem 4.3-3):

$$\lambda = -N \quad \text{and} \quad P(k) = \sum_n f(k|n)/N \tag{4.3-14}$$

where N is the total number of pixels, $N = \Sigma_n 1$. Because $P(k)$ is a relative number of pixels belonging to object k (or originating from source k),

$$N_k = \sum_n f(k|n) \tag{4.3-15}$$

is a number of pixels on the object k, which is intuitively an appealing interpretation.

Expression (4.3-13) for the fuzzy class memberships is the famous Bayesian expression for the a posteriori probability of class k. It is called *a posteriori* because it is computed *after* the datum $\mathbf{X}_n$ observation. As mentioned in Chapter 1, this Bayes expression was, probably, the first mathematical tool of combining a priori knowledge with data in presence of uncertainty. In our case, (4.3-13) can be interpreted as a probability only after the convergence of the MLANS estimation process. Before MLANS converges and parameters are estimated, these quantities contain additional uncertainty due to unknown values of parameters $\mathbf{S}_k$, and, thus, cannot be considered as probabilities; instead, they can be considered as fuzzy memberships.

Similar to Eq. (4.2-4), a dynamic equation can be used in place of Eq. (4.3-13),

$$df(k|n)/dt = f(k|n) \sum_{k'} \left[\delta_{kk'} - f(k'|n)\right] \, \mathrm{ll}'(n|k') \, \mathbf{M}'_{k'} \cdot d\mathbf{S}_{k'}/dt \tag{4.3-16}$$

where

$$\mathrm{ll}'(n|k') = \partial \, \ln \, \mathrm{pdf}\,(\mathbf{X}_n|H_{k'}) \,/\partial \mathbf{M}_{k'}; \quad \mathbf{M}'_{k'} = \partial \mathbf{M}_{k'}/\partial \mathbf{S}_{k'} \tag{4.3-17}$$

If the Gaussian densities are used for conditional pdfs, Eq. (4.3-6), the above equations can be written more explicitly. To simplify the following derivation, we consider covariances $\mathbf{C}_k$ in Eq. (4.3-6) as known and only expected values $\mathbf{M}_k$ to be functions of model parameters $\mathbf{S}_k$. In this case (see Problem 4.3-1 for further details),

$$\mathrm{ll}'(n|k') = \partial \, \ln \, \mathrm{pdf}\,(\mathbf{X}_n|H_{k'}) \,/\partial \mathbf{M}_{k'} = (\mathbf{X}_n - \mathbf{M}_{k'})^T \, \mathbf{C}_{k'}^{-1} \tag{4.3-18}$$

and the MLANS learning equations are as follows,

$$d\mathbf{S}_k/dt = \sum_n f(k|n) \, \left[(\mathbf{X}_n - \mathbf{M}_k)^T \, \mathbf{C}_k^{-1}\right] \mathbf{M}'_k \tag{4.3-19}$$

$$df(k|n)/dt = f(k|n) \sum_{k'} \left[\delta_{kk'} - f(k'|n)\right] \, \left[(\mathbf{X}_n - \mathbf{M}_{k'})^T \, \mathbf{C}_{k'}^{-1}\right] \mathbf{M}'_{k'} \cdot d\mathbf{S}_{k'}/dt \tag{4.3-20}$$

Specifying all the vector and matrix indexes explicitly, these equations are written as

$$dS_k^a/dt = \sum_n f(k|n) \, \left[\left(X_{i,n} - M_{ik}\right) C_{ijk}^{-1} M_{jk}^{;a}\right] \tag{4.3-21}$$

$$df(k|n)/dt = f(k|n) \sum_{k'} \left[\delta_{kk'} - f(k'|n)\right] \left[\left(X_{i,n} - M_{ik'}\right) C_{ijk'}^{-1} M_{jk'}^{;a}\right] \cdot dS_{k'}^a/dt \tag{4.3-22}$$

here, index a refers to the components of the vector of model parameters, and indexes i, j refer to the components of the data vectors; summation is assumed over repeated indexes

a, i, j; and (;) denotes partial derivatives with respect to parameters S with the corresponding index:

$$M_{ik}^{;a} \equiv \partial M_{ik} / \partial S_k^a \tag{4.3-23}$$

Equations (4.3-19) through (4.3-22) have a property that could be undesirable: their stationary points ($d\mathbf{S}_k/dt = 0$) include the points where the model is insensitive to parameter variations, $\partial \mathbf{M}_k / \partial \mathbf{S}_k = 0$, even if the model does not match the data. This could be mitigated by the following modification,

$$dS_k^a/dt = \left[\sum_n f(k|n)\ M_{ik}^{;a}\ C_{ijk}^{-1} M_{jk}^{;b} \right]^{-1} \left[\sum_n f(k|n)\ \left(x_{i,n} - M_{ik}\right) C_{ijk}^{-1} M_{jk}^{;b} \right] \tag{4.3-24}$$

where a summation is assumed over repeated indexes b, i, j. Here, the denominator becomes small near the points of insensitivity, and a stationary point ($d\mathbf{S}_k/dt = 0$) is attained only if the model, on average, matches the data. An appropriateness of the normalization of Eq. (4.3-24) is seen also from the fact that the iteration increment factor dt is dimensionless. Equation (4.3-24) can be viewed as a modification of the Newton gradient method, which generally leads to a faster convergence. (Issues of this paragraph are further addressed in Problems 4.3-6 through 4.3-8.) To summarize, MLANS learning equations (4.3-19, 4.3-21 or 4.3-24) and (4.3-13, 4.3-20, or 4.3-22) define a convergent dynamic system. A proof of convergence is considered in Section 4.6.

Fuzziness of MFT agent-classes is controlled by the covariance matrixes, $\mathbf{C}_k$; a large covariance corresponds to a high degree of fuzziness and vice versa. During learning, fuzziness should be matched to the parameter uncertainty. This can be achieved by the following equation, combining estimated and controlled covariances:

$$\mathbf{C}_k = \mathbf{C}_{0,k} \exp(-ct) + \sum_n f(k|n) \left[(\mathbf{X}_n - \mathbf{M}_k) (\mathbf{X}_n - \mathbf{M}_k)^T \right] / \sum_n f(k|n) \tag{4.3-25}$$

Here, $\mathbf{C}_{0,k}$ is a large covariance corresponding to the initial uncertainty of the model $\mathbf{M}_k$. During learning, the internal MFT time parameter t grows and the contribution of this fuzzy uncertainty into the covariance $\mathbf{C}_k$ diminishes. Therefore, $\mathbf{C}_k$ is gradually reduced to the second item, which gives the ML covariance estimation (see Problem 4.3-9). Parameter c in the exponent controls the speed of learning and convergence: convergence can be attained only when the first item is much smaller than the second one. Selection of the c value is a matter of tradeoff: if c is large, the convergence is fast, however, there is a chance that the covariance will be reduced to a small value prematurely, before the parameters reach their true values (or the best estimated values corresponding to the global maximum of the likelihood). If c is small, convergence is long, but there is a better chance of attaining accurate parameter values in complicated cases. Not only the MFT agents, but also living beings face this dilemma: some creatures mature fast and cannot adapt thereafter. A similar effect exists on a much shorter time scale within a single act of recognition: before an object or concept is recognized, the fuzziness of its internal agent-model should correspond to the uncertainty about what is being recognized. This is the mathematical expression of Aristotelian Forms-as-potentialities that evolve into concepts in the process of learning.

4.4 Shannon–Einsteinian MFT

Whereas Bayesian likelihood similarity measures a degree of stochastic deviations of data from its mean value given by the model, a different concept of similarity is introduced here that measures the amount of information in the model about the world. Information was defined by Shannon as a measure of certainty of choice among the finite number of alternatives. Thus, to define an information measure, we need to define alternative states of the world and compute their number. Quantum physics provides the definition of the states of any system and rules for their computation. But before turning to quantum theory, we begin with a simple idea due to Einstein about the nature of the electromagnetic spectrum, which precedes quantum theory and which lets us relate information to likelihood and derive information-based similarity in a simple and straightforward way. This similarity measure is extended to vision in general, since the visual system senses electromagnetic spectra; and we mention its applicability to acoustic signals as well. Then we discuss basic definitions of information theory and related issues of quantifying states of the world. We conclude this section with a brief overview of the history of and relationships among various methods based on information and entropy maximization.

4.4.1 Einstein, Likelihood, and Electromagnetic Spectrum

Einstein interpreted the electromagnetic spectrum as a probability density function (pdf) of photon frequency (Einstein and Hopf, 1910). This is a different type of pdf than usually considered in statistical estimation. Because the concept of pdf is a basis for statistical estimation, likelihood, and information measures, "Einsteinian" pdf leads to different results than the classical ML estimation considered in the previous section. A specific point of difference is the attribution of randomness. Usually, in statistical estimation, randomness is attributed to measured quantities as follows: if an image is produced by light intensity, an intensity of the pixel is considered as a random quantity and pdf models are designed to model this randomness. Or, if a radar measures Doppler spectra, a signal intensity or power in a Doppler cell (a sample) is considered a random quantity, and pdf models are constructed accordingly (say, for an image pixel intensity it could be a Gaussian density, for the spectrum it could be a χ^2 density, etc.). Contrary to this, Einstein interpreted the spectrum as a pdf of photon frequency. That is, the Doppler cell frequency is to be considered random, rather than the signal intensity or power. A similar interpretation is valid for phonons of acoustic spectra (speech, seismic signals, etc.) and for any signal field obeying Bose–Einstein statistics (bosons); and it could be extended to any classical field, when quantum effects are not important.

At first, the Einsteinian idea may seem bizarre: a radar measures signals at a set of frequency values predetermined by the radar operational parameters. However, consider a measurement process as detecting individual photons. Every classical or quantum measurement of electromagnetic fields can be described in this way. For an individual photon, there is no "intensity" but only frequency and polarization. Therefore, from the first principles of physics, the frequency can be a random quantity.

According to this idea, the models of pdf, to be developed in this section, are the models of spectra. To emphasize the difference between conventional statistical pdf and Einsteinian

pdf, we denote pdf in this section by $F(\omega)$. The Einsteinian interpretation requires proper normalization; since spectrum $S(\omega)$ is measured in units of energy, its interpretation as a pdf requires normalization on a photon energy in order to obtain a measure of the number of photons. A photon energy ε is related to its frequency ω,

$$\varepsilon = \hbar\omega \tag{4.4-1}$$

where $\hbar$ is the Plank constant. Therefore, the number of measured photons

$$N_\omega = S(\omega)/\hbar\omega \tag{4.4-2}$$

and $F(\omega)/\hbar\omega$ is a model pdf for a single photon with frequency ω, normalized in a standard way

$$\sum_\omega F(\omega)/\hbar\omega = 1 \tag{4.4-3}$$

In macroscopic systems, photons are statistically uncorrelated (most often, this is also true for microscopic systems as well).[4] Therefore, for an ensemble of photons $n = 1, \ldots, N$, the joint pdf or likelihood L is a product over individual photons

$$L = \prod_n F(\omega_n)/\hbar\omega_n = \prod_\omega [F(\omega)/\hbar\omega]^{N_\omega} = \prod_\omega [F(\omega)/\hbar\omega]^{S(\omega)/\hbar\omega} \tag{4.4-4}$$

The second equation here is obtained as follows. The product over individual photons, n, is split into two terms: a product over photons with a fixed frequency ω and a product over various frequencies ω. There are N_ω photons with a fixed frequency ω, all distributed according to the identical pdf models $[F(\omega)/\hbar\omega]$; this leads to the above equation.

As in previous sections, the overall model $F(\omega)$ is composed of multiple agents or submodels $F(\omega|k)$, modeling alternative signal sources, or objects of recognition,

$$F(\omega) = \sum_k F(\omega|k), \quad k = 1, \ldots, K \tag{4.4-5}$$

Often, signals can be considered as being produced by incoherent contributions from several sources; the above equation then models signals according to the first principles, otherwise it is an approximation (several $F(\omega|k)$-models might be needed to model individual sources). Combining the above equations we obtain the Einsteinian likelihood or AZ-similarity as

$$\begin{aligned} \text{AZ-}L &= \max_{\mathbf{S}_k} \left\{ \prod_\omega \left[\sum_k F(\omega|k)/\hbar\omega \right]^{S(\omega)/\hbar\omega} \right\} \quad \text{or} \\ \text{AZ-}LL &= \max_{\mathbf{S}_k} \left\{ \sum_\omega [S(\omega)/\hbar\omega] \ \ln \left[\sum_k F(\omega|k)/\hbar\omega \right] \right\} \quad \Rightarrow \{\mathbf{S}_k\} \end{aligned} \tag{4.4-6}$$

Maximization of the Einsteinian likelihood [Eq. (4.4-6)] is achieved in a manner similar to the corresponding equations of the general MFT and MLANS. Again, we have to account

for the normalization constraint (4.4-3), which can be done by the Lagrangian multiplier method. Using the gradient ascent, similar to Eqs. (4.2-1) through (4.2-4), we obtain

$$d\mathbf{S}_k/dt = \sum_{\omega} [S(\omega)/\hbar\omega]\ f(k|\omega)\ [\partial \ln(F(\omega|k)/\partial \mathbf{S}_k] + \lambda\,(\partial/\partial \mathbf{S}_k)\left[\sum_{\omega} F(\omega)/\hbar\omega - 1\right] \tag{4.4-7}$$

Here, the fuzzy membership of the frequency ω in concept-model k or, alternatively, partial activations of concept-agent k by ω, $f(k|\omega)$, are determined by partial similarities, as in Section 4.1,

$$f(k|\omega) = F(\omega|k)/F(\omega) \tag{4.4-8}$$

The following sections will relate the Einsteinian likelihood to the entropy of the photon ensemble and to Shannon's information, but first, let us consider specific types of spectrum models related to Gaussian mixtures used for MLANS.

4.4.2 Einsteinian Gaussian Mixture Model

Specific shapes of the parametric models for $F(\omega|k)$ can be determined based on the physics or phenomenology of the process under consideration, or a general type of flexible parametric model can be selected. Here we consider a general type model suitable for modeling a variety of signals as a superposition of Gaussian functions. Superposition models of the type (4.4-5) are called in statistics mixture models, and we call each mixture component $F(\omega|k)$ a conditional model, submodel, or agent-model corresponding to source k. According to our normalization (4.4-3), the Gaussian agent-model is given by

$$\begin{aligned} F(\omega|k) &= \hbar\omega\ A_k\ G(\omega|k), \quad k = 1, \ldots, K \\ G(\omega|k) &= (2\pi)^{-1/2}\,(\sigma_k)^{-1}\ \exp\left\{-0.5\ (\omega - \omega_k)^2/\sigma_k^2\right\} \end{aligned} \tag{4.4-9}$$

Here, A_k is a submodel amplitude, ω_k is the submodel mean frequency, and σ_k is the submodel frequency standard deviation. A multiplicative term $\hbar\omega$ is introduced according to our normalization, so that $[A_k\ G(\omega|k)]$ is measured in units of a photon number, and $G(\omega|k)$ is interpreted as the conditional pdf of photons from source k.

Because of the $\hbar\omega$ factor, the above model is different from Gaussian mixture models considered previously, and we call it an Einsteinian Gaussian mixture. These models do not form a complete set of basis functions, since all $F(\omega|k)$ are 0 at $\omega = 0$. This is related to the well-known "infrared catastrophe" in quantum electrodynamics: the number of photons might go to infinity at zero frequency. This would prevent normalization of the pdf, and, in order to maintain the Einsteinian interpretation of the spectrum as a pdf, the power at zero frequency should be zero. However, in most practical cases this does not represent a modeling problem: most electromagnetic signals are carried by a high-frequency carrier field, so changes in frequency are very small in relative terms, the $\hbar\omega$ factor is nearly constant, so that $F(\omega|k)$ form a complete set of functions and any function N_ω can be modeled using the above models. When low frequencies are included in the data,

as might be the case with acoustic signals, the actual zero frequency is never measured and the numerical problem at zero frequency is avoided by using mid-sample frequency for a sampling interval including zero. Therefore, the Einsteinian Gaussian mixture model is a general type of model that can be used for a variety of signals.

The normalization constraint (4.4-3) affects only the amplitude parameters, A_k, of the Einsteinian Gaussian mixture model. This can be seen as follows. Substitute (4.4-9) into (4.4-3):

$$\sum_{\omega}\sum_{k} A_k G(\omega|k) = 1 \tag{4.4-10}$$

For simplicity of notations, let us use a frequency sampling interval as the unit of frequency, $\Delta\omega = 1$, then

$$\sum_{\omega} G(\omega|k) \approx \int G(\omega|k)\, d\omega \equiv 1 \tag{4.4-11}$$

By exchanging the order of summation in (4.4-10) and combining with (4.4-11), the constraint can be written as

$$\sum_{k} A_k = 1 \tag{4.4-12}$$

The approximation in considering the discrete Gaussian densities to be normalized (4.4-11) is accurate within a few percent for $G(\omega|k)$-models that are wider than few samples ($\sigma_k > 1$ in units of sample numbers), so the approximation is accurate.

The above model is characterized by three parameters per Gaussian submodel: the amplitude, the mean, and the standard deviation. The ML estimation equations for these parameters are derived from the general MFT Eq. (4.4-7). In case of Gaussian models, they can be simplified as discussed in Problem 4.4-1. These equations can also be derived by using the Estimation-Maximization algorithm (EM); because EM is a powerful and useful tool, we consider it in Problem 4.4-3. The result is the following MFT equations[5]:

$$A_k = N_k/N, \quad N_k = \sum_{\omega} f(k|\omega)\, N_\omega, \quad N = \sum_{\omega} N_\omega, \quad N_\omega = S(\omega)/\hbar\omega \tag{4.4-13}$$

$$\omega_k = \sum_{\omega} f(k|\omega)\, N_\omega\, \omega/N_k \tag{4.4-14}$$

$$\sigma_k^2 = \sum_{\omega} f(k|\omega)\, N_\omega\, (\omega - \omega_k)^2 / N_k \tag{4.4-15}$$

In these equations, $f(k|\omega)$ in the right-hand side are evaluated using parameters estimated at the previous iteration, and the left-hand side yields the current-iteration parameter value. The new parameter values are used in the right-hand side for the next iteration, etc. In this way, the equations define an iterative system: beginning with some values of parameters, the agent-submodels $F(\omega|k)$ are computed according to Eqs. (4.4-9), followed by the computation of $f(k|\omega)$ according to (4.4-8); on the next iterations, the parameter values are recomputed according to Eqs. (4.4-13), (4.4-14), (4.4-15), etc., until convergence. The convergence is determined by requiring that parameter changes are small from iteration to iteration. The

convergence is always attained, as proved in Section 4.6. On convergence of the estimation process, $f(k|\omega)$ can be interpreted as the a posteriori Bayes probability that a photon at frequency ω has originated from the source (or submodel) k. Correspondingly, N_k is the number of photons from the source k, and N is the total number of photons.

4.4.3 Equilibrium of the Photon Ensemble

This section relates the Einsteinian likelihood to the entropy of the photon ensemble. We will see that the ML Eqs. (4.4-7) and (4.4-8) describe an equilibrium state of a "universe" consisting of the estimation system and the outer world, containing the physical ensemble of photons. In the process of equilibration, the estimation system models itself and the outer world in such a way that the observed photon ensemble is in equilibrium. An equilibrium state of a physical system is determined by its entropy. Thus, we need to compute the entropy of the "universe," modeled by the estimation system. In this section we consider an estimation system whose internal entropy is not a function of the model parameters, so we need only consider the estimated entropy of the photon ensemble.

Entropy of a physical system is related to its physical states. The expected number of observations of a photon is proportional to the number of the photon physical states compatible with the observations. And the expected number of observations is proportional to a pdf. Therefore, according to Einstein's interpretation, a spectrum model $F(\omega)$ is proportional to a number of physical states, Φ_ω, for a single photon at each frequency,

$$F(\omega) = \text{const} \cdot \hbar\omega \cdot \Phi_\omega \tag{4.4-16}$$

The equations of physical equilibrium can be derived by using a standard textbook procedure, which is now briefly described. Since our estimation procedure deals with fixed observation data, the total number and energy of photons are fixed. In statistical physics, a system with fixed energy and number of particles is called a canonical ensemble. The equilibrium of a canonical ensemble is obtained by maximizing the entropy of the ensemble, E, subject to the constraints of conservation of energy ε and photon number N. According to Eqs. (4.4-1), (4.4-2), and (4.4-3), this results in the following constraints on our model $F(\omega)$,

$$\varepsilon = \sum_\omega S(\omega) = N \sum_\omega F(\omega); \quad N = \sum_\omega S(\omega)/\hbar\omega = N \sum_\omega F(\omega)/\hbar\omega \tag{4.4-17}$$

The ensemble entropy, E, is a logarithm of the number of states available to the system, Γ (gamma). It is a product of the number of states available at each frequency, Γ_ω, and it is computed as follows.[6] Of N_ω photons with frequency ω, every one can be in any of Φ_ω states, giving $(\Phi_\omega)^{N\omega}$ combinations. Photons with the same frequency are indistinguishable, therefore, permutations among photons correspond to the same states. The number of permutations is $(N_\omega!)$. Thus, the number of states for N_ω photons, Γ_ω is

$$\Gamma_\omega = (\Phi_\omega)^{N\omega} / (N_\omega!) \tag{4.4-18}$$

where the numerator is the total number of all combinations and the denominator accounts for the permutations of the equivalent photons. Taking the logarithm of the above, and approximating $\ln(N!) \approx N \ln N$, we obtain the entropy of the photon ensemble,

$$E = \ln \Gamma = \ln \prod_{\omega} \Gamma_{\omega} \approx \sum_{\omega} N_{\omega} \ln [\Phi_{\omega}/N_{\omega}] \tag{4.4-19}$$

Maximization of this expression subject to constraints (4.4-17) is considered in Problem 4.4-4. In particular, it is shown in Problem 4.4-4 that the constraint on the number of particles is equivalent to the normalization of the model $F(\omega)$ in the previous section. For a fairly broad set of conditions, entropy maximization is equivalent to Eqs. (4.4-7) and (4.4-8). Thus, the ML estimation using "Einsteinian likelihood" is equivalent to finding an equilibrium of the photon ensemble.

4.4.4 Einsteinian Likelihood and Shannon's Mutual Information

The Einsteinian likelihood will be related now to information. This section summarizes the relationship; the next three sections will review the fundamental concepts of information theory and derive the relationship stated here.

In his classical work published in 1948 in the Bell Laboratory journal, Shannon introduced a concept of information and a related concept of the mutual information contained in the receiver (received message) about the source (sent message). For our purpose, this concept can be formulated as follows. We identify the source with the measured data $S(\omega)$ and the receiver with the internal model, $F(\omega)$. Then, the mutual information in the model about the data is given by

$$I = \sum_{\omega} [S(\omega)/\hbar\omega] \ln [F(\omega)/\hbar\omega] \tag{4.4-20}$$

This expression is identical to the Einsteinian likelihood (4.4-6); for estimation purposes, it is equivalent to the entropy (4.4-19), and maximization of any of these expressions leads to the same set of parameter values. The next section reviews Shannon's definition of information and arguments of why (4.4-20) is a measure of mutual information. We will also consider in detail a potential source of confusion: if entropy and information are "opposite" quantities,[7] how come the maximization of information is equivalent to maximization of entropy? The answer is in that the entropy (4.4-19) is defined for a different system than information (4.4-20). In the process of learning, the entropy of the "universe" consisting of the data and internal model increases, while the total amount of information contained in the universe decreases. This decrease of information in the universe corresponds to the correlation between states of the internal model and the data and to an increase of the mutual information, that is to learning.

4.4.5 Information and Alternative Choice States

Information, as defined by Shannon, is a measure of certainty of choice among a finite number of alternatives. Thus, to define an information measure, it is necessary to define a number of alternative choice states. Consider the universe (U) as consisting of two systems: one, an intelligent system (IS), and the other, the outer world (W). Denote the data about the world available to IS as X (in the previous sections, the data are the set of spectral values, $X = \{S(\omega)\}$). We are interested in the best modeling or representation of X within IS, and we would like to develop a measure of "the best" based on information in IS about

X. Therefore, we need to know the numbers of state in each system. The number of states we denote by Γ. We use indexes to denote particular systems: Γ_W and Γ_{IS} are the total numbers of states in W and in IS. Data X, in general, does not specify the state of the world unambiguously; the number of states in W compatible with data X we denote as $\Gamma_{W|X}$ (read: gamma W given X).

Information and entropy are a pair of opposites (see Note 7). Entropy (E) is a measure of uncertainty and is given by a logarithm of the number of choices. Thus, the amount of information needed to identify the state of the world without uncertainty is $I_W = \ln\ \Gamma_W$, alternatively, $E_W = \ln\ \Gamma_W$ is a measure of the maximal uncertainty about the state of the world. (Being concerned with coding and transmission of information, Shannon would also say that $\ln\ \Gamma_W$ is the total amount of information that could be "recorded" in W.) When discussing entropy and information, we must be careful about which states and systems are considered: I_W is information about the world contained in the unambiguous knowledge of the world state, whereas E_W is uncertainty about the world state when nothing is known about it. The amount of uncertainty about W, when X is known, is $E_{W|X} = \ln \Gamma_{W|X}$. The amount of information contained in X about the state of the world is measured by a reduction of uncertainty about W due to X:

$$I_{W|X} = E_W - E_{W|X} \tag{4.4-21}$$

Thus, $I_{W|X}$ and $E_{W|X}$ are opposite quantities: an increase of one is equal to a decrease in the other.

If the system and world are not interacting, the states of the system are independent of the states of the world, the total number of states in the universe is a product of two terms, $\Gamma_{U,0} = \Gamma_{IS} \cdot \Gamma_W$. And the total amount of information needed to identify unambiguously a state of the universe is a sum, $I_{U,0} = \ln\ \Gamma_{U,0} = \ln\ \Gamma_{IS} + \ln\ \Gamma_W$. But, if there is no interaction, there is no learning, and since we are interested in learning, we would like to consider a case, when IS and W interact. In the case of interacting IS and W, states of IS are to some extent determined by states of W and vice versa, and the total amount of information needed to describe U unambiguously is reduced, $I_U \le I_{U,0}$. This reduction of the total information in U is due to an increase of information in IS about W, which is learning. Thus, the measure of the information in IS about the world is

$$I_{W|IS} = I_{U,0} - I_U \tag{4.4-22}$$

This information gives a measure of similarity between the world W (as represented by the available data X) and the intelligent system IS (as contained in the IS' internal model or representation of the world). To define this information-based similarity between the model and the data, we need to compute the numbers of states for the considered systems.

Computation of the numbers of states of various systems depends on the goal: which variations in system parameters are of interest or importance and should be counted as different states, and which should be ignored? Such a general formulation could lead to a definition of information contingent on *meaning* of various states and on defining *intelligence*. Of course, understanding of intelligence is an ultimate goal of our study, however, a complete solution of this problem is still far away, if ever achievable. To progress toward understanding of intelligence, we need simpler basic definitions that are not contingent on the solution of the entire problem. And should not scientists assume

that the evolution of natural intelligence has to be possible before the final goal is reached (whichever this goal might be)?

Shannon approached this conundrum by suggesting that an engineering problem should be limited to counting the numbers of well-defined states, such as letters in the alphabet, or decisions in a control system. The problem of *defining* "choice" states is supposed to be solved by general culture (alphabet) or by an engineering design based on human intellectual analysis. Thus, the definition of information is not related to *the meaning* that could be eventually associated with states in the process of learning, but is rather related to probabilities of predefined states and, ultimately, to a notion of equiprobable outcomes, which has been at the foundation of the probability theory since its inception.

This contingency of the information measure on the definition of choice states can be illustrated by considering information contained in a piece of text. If alphabet characters are used as choice states, the information is determined by a set of used characters and their probabilities in the language; this might be a useful measure for a teletype device transmitting text character by character. If individual words are used as choice states, the information is determined by a set of used words and their probabilities in the language; this might be a useful measure for a recognition device that recognizes words one by one. The two measures of information will be very different, and neither is related to the information extracted from the text by a human reader, whose information measure is related to personal mental states.

In our case of modeling "the world," the step of defining states is naturally accomplished due to the quantum nature of the world: according to principles of quantum physics, any physical system is characterized by a finite number of quantum states and techniques for counting these states are well developed in quantum statistical physics. For microscopic systems, including various microscopic imagery devices, counting the actual number of quantum states emitting the detected photons could be adequate. For macroscopic systems, such as photons emitted by a part of a scene that are measured by a detector in a CCD camera, the actual number of quantum states is usually unknown and difficult to model from the first principles. For these problems, "counting states" is to a certain extent a metaphor, instead of actual quantum states, we will use approximate models, and success of such modeling depends on adequacy of the models. As is usually the case, making these models flexible and adaptive to the data is an essential step toward robustness of the modeling procedure.

To summarize this discussion, the fundamental limitation of the concept of information is its contingency on the definition of "choice" states. Because the states of the world of importance to us are those that affect our internal mental states, emotions, and behavior, their definition is contingent on the solution of the entire problem of *intelligence*. Information theory does not attempt to solve this problem. Nevertheless, the concept of information is one of the most fundamental and useful notions developed in this century, and the information-based measure of similarity that we develop below is a fundamental and useful measure. And could it be that our "freedom of choice" is related to our ability to learn about the world according to information measures, which are "objective" and independent of our ultimate "subjective" choices?

4.4.6 Mutual Model-Data Information

To count numbers of states using quantum concepts we would not need any knowledge of quantum theory, a few necessary concepts and terminology are introduced as needed,

and the procedure is essentially the same as we have already encountered in Section 4.4.3. A standard terminology used in physics is to call a state of the world described by X a *macrostate*, to distinguish it from a quantum or *microstate* that gives a complete exhaustive description of the physical state of a system. According to the first principles of statistical physics, every microstate accessible to a statistical system in a state of equilibrium is equiprobable. Thus, we have approached the problem of defining "choice" states by taking a "physicist's" point of view: our internal model ought to represent the world "as it actually is" according to the first laws of physics.

If one considers an "ensemble of the worlds," the proportion of the worlds in a macrostate X is simply $\Gamma_{W|X}/\Gamma_W$. Equivalently, $\Gamma_{W|X}/\Gamma_W$ is the probability of finding the world in a macrostate X. In case of X having continuous values, $\Gamma_{W|X}$ is a number of states corresponding to an interval between X and $X + dX$, and

$$\text{pdf}(X) = \Gamma_{W|X}/\Gamma_W \tag{4.4-23}$$

Along with the definitions of entropy and information, this establishes relationships between pdf and entropy or information. Using (4.4-21), we obtain the information about the world contained in data X as a reduction of uncertainty from E_W to $E_{W|X}$,

$$I_{W|X} = E_W - E_{W|X} = -\ln\left(\Gamma_{W|X}/\Gamma_W\right) = -\ln\ \text{pdf}(X) \tag{4.4-24}$$

To analyze change of information and entropy in the process of learning, we need to consider states of the "universe" consisting of the intelligent system and the world. In the process of learning, internal variables of the intelligent system, Y, become correlated with the data about the world, X, and the total number of states in the universe is reduced from $\Gamma_{U,0}$ to $\Gamma_{U|X,Y}$. The joint density $\text{pdf}(X, Y)$ is determined by the number of states of the universe consisting of IS and W, similar to (4.4-23),

$$\text{pdf}(X, Y) = \Gamma_{U|X,Y}/\Gamma_{U,0} \tag{4.4-25}$$

And, the total information that is given about the universe by (X, Y) is the reduction of uncertainty from $E_{U,0}$ to $E_{U|X,Y}$:

$$\begin{aligned} I_{U|X,Y} &= E_{U,0} - E_{U|X,Y} = \ln\Gamma_{U,0} - \ln\Gamma_{U|X,Y} = -\ln\left[\Gamma_{U|X,Y}/\Gamma_{U,0}\right] \\ &= -\ln\ \text{pdf}(X, Y) \end{aligned} \tag{4.4-26}$$

Using the rule of conditional probabilities, $\text{pdf}(X, Y) = \text{pdf}(X|Y)\text{pdf}(Y)$, and combining Eqs. (4.4-23) through (4.4-26), we obtain (see Problem 4.4-2),

$$\begin{aligned} I_{U|X,Y} &= -\ln\ \text{pdf}(X, Y) = -\ln\{[\text{pdf}(X)\text{pdf}(Y)][\text{pdf}(X|Y)/\text{pdf}(X)]\} \\ &= I_{U,0|X,Y} - \ln[\text{pdf}(X|Y)/\text{pdf}(X)] \end{aligned} \tag{4.4-27}$$

Comparing this to (4.4-22), the mutual information in IS about the world is given by

$$I_{W|IS} = I_{U,0} - I_U = \ln[\text{pdf}(X|Y)/\text{pdf}(X)] = \ln\ \text{pdf}(X|Y) - \ln\ \text{pdf}(X) \tag{4.4-28}$$

The mutual information is defined here according to the discussion in the previous section, which we repeat again because of its importance: the total information that could be stored

in the universe is reduced by the amount of the mutual information in the intelligent system about the world. The amount of information that (X, Y) gives about the universe depends on the total amount that could be stored. The universe stores more information when IS and W are independent and X and Y are not correlated, because when X and Y are correlated, some of the information (namely, mutual information in Y about X) is redundant. Thus, when X and Y are correlated, the information is reduced by $I_{\mathrm{W|IS}}$. It follows, that $I_{\mathrm{W|IS}}$ is a positive quantity. When the intelligent system learns, the mutual information $I_{\mathrm{W|IS}}$ increases.

We define Shannon's similarity as mutual information MI between the world and the intelligent system, or to be more specific, between the data X and the internal model or internal representation of the world within the intelligent system,

$$LL_{\mathrm{Shannon}} = I_{\mathrm{W|IS}} \tag{4.4-29}$$

According to (4.4-28), Shannon's similarity consists of two items. The first one, ln pdf $(X|Y)$, depends on the internal model and it parameters, Y. The second item ln pdf(X) depends on the unknown number of states of the world, but it does not depend on the model parameters and therefore does not affect the estimation procedure. Therefore, for the estimation purpose, it is sufficient to consider only the first item. This first item is related to the Einsteinian likelihood in the next section.

4.4.7 Shannon–Einsteinian Similarity

Here we obtain expressions for the numbers of states that determine mutual information and relate it to the Einsteinian likelihood derived in Sections 4.4.1 and 4.4.2. The Einsteinian interpretation of spectrum does not have to be limited to the frequency-domain spectra, but is naturally extended to intensity or power densities in any coordinates, e.g., two-dimensional domains of time-frequency spectra, regular angle–angle intensity imagery, or higher dimensional domains such as time–frequency–range–angle imagery. The extension is a straightforward one. Let us denote the general multidimensional image coordinates as $\mathbf{x}$, image intensity data, $S(\mathbf{x})$, and the model of the number of states, $\Phi(\mathbf{S}, \mathbf{x})$, where $\mathbf{S}$ are model parameters. The model $\Phi(\mathbf{S}, \mathbf{x})$ gives the number of microstates for a single photon in a macrostate $\mathbf{x}$. And, the number of photons in pixel $\mathbf{x}$ is

$$N_{\mathbf{x}} = S(\mathbf{x})/\hbar\omega_{\mathbf{x}} \tag{4.4-30}$$

where $\omega_{\mathbf{x}}$ is a frequency of pixel $\mathbf{x}$; it may be included in $\mathbf{x}$ as for time-frequency data, $\mathbf{x} = (t, \omega)$, or multiband (color) imagery, $\mathbf{x} = (x, y, \omega)$, or $\omega_{\mathbf{x}}$ may be a constant, playing no important role as in black–white imagery, $\mathbf{x} = (x, y)$. Image intensity (or signal power) is given by $S(\mathbf{x})$, so the world data $X = \{S(\mathbf{x})\}$ and the internal variables of the intelligent system specify the model and model parameters, $Y = \{\Phi(\mathbf{S}, \mathbf{x}), \mathbf{S}\}$. For acoustic signals we should be talking about phonons rather then photons, but this does not alter the general argument presented here, which is applicable to various types of signal or image data. The Shannon's similarity part that depends on the model parameters is given by ln pdf$(X|Y)$, which is computed by counting the number of states of the world given the model,

$$\mathrm{pdf}(X|Y) = \Gamma_{\mathrm{W}|X,Y} / \Gamma_{\mathrm{W}|Y} \tag{4.4-31}$$

The numerator here is the number of states of the world, given the data and model. Its computation is similar to that in Section 4.4.3 (see Problem 4.4-3),

$$\ln\ \Gamma_{W|X,Y} = \sum_{\mathbf{x}} N_{\mathbf{x}}\ \ln\left[\Phi(\mathbf{S},\mathbf{x})/N_{\mathbf{x}}\right] \tag{4.4-32}$$

or, if the photons in $\mathbf{x}$-macrostate are not equivalent,[8]

$$\ln\ \Gamma_{W|X,Y} = \sum_{\mathbf{x}} N_{\mathbf{x}}\ \ln \Phi(\mathbf{S},\mathbf{x}) \tag{4.4-33}$$

From previous derivations, we know that the above two expressions lead to the same estimation equations (because the only difference is an item $\sum\ N_{\mathbf{x}}\ \ln\ N_{\mathbf{x}}$, which is not a function of parameters $\mathbf{S}$). We would prefer to use this latter expression because it is a little simpler.

The denominator in (4.4-31) accounts for the a priori information. The a priori world state could be specified only by the invariant properties (that are not changed during estimation). If nothing is known about the state of the world a priori, the denominator does not affect the estimation process. This type of unconstrained estimation might be unsatisfactory, e.g., if the model $\Phi(\mathbf{S},\mathbf{x})$ may reach infinite values (we know a priori that this is not appropriate). In our original Einsteinian model in Section 4.4.1, this was prevented by a proper choice of the $\Phi(\mathbf{S},\mathbf{x})$ model (namely, that the model is normalizable, independent of the parameter values). As we have seen in Section 4.4.2, this was equivalent to the conservation of the number of particles during the estimation process, or in other words, the number of particles constituted the a priori information. To repeat again: for parameter estimation purposes, a constraint on the model,

$$\Phi = \sum_{\mathbf{x}} \Phi(\mathbf{S},\mathbf{x}) \tag{4.4-34}$$

is equivalent to a constraint on the number of particles (particle conservation),

$$N = \sum_{\mathbf{x}} N_{\mathbf{x}} \tag{4.4-35}$$

So it would not surprise us that using the a priori number of states of the world corresponding to the above constraints, $\Gamma_{W|Y} = \Gamma_{W|N,\Phi}$, we obtain the Einsteinian likelihood from the above equations. The total number of states of N particles distributed among Φ single-particle states is

$$\Gamma_{W|N,\Phi} = \Phi^N \quad \text{or} \quad \ln\ \Gamma_{W|N,\Phi} = N\ \ln\ \Phi = \sum_{\mathbf{x}} N_{\mathbf{x}}\ \ln\ \Phi \tag{4.4-36}$$

Combining the above with (4.4-31) and (4.4-33), we obtain

$$\begin{aligned} \ln\ \mathrm{pdf}(X|Y) = \ln\ \Gamma_{W|XY} - \ln\ \Gamma_{W|N,\Phi} &= \sum_{\mathbf{x}} N_{\mathbf{x}}[\ln\ \Phi(\mathbf{S},\mathbf{x}) - \ln\ \Phi] \\ &= \sum_{\mathbf{x}} N_{\mathbf{x}}\ \ln[\Phi(\mathbf{S},\mathbf{x})/\Phi] \end{aligned} \tag{4.4-37}$$

Due to (4.4-34), the term in brackets is normalized like a pdf,

$$\sum_{\mathbf{x}} [\Phi(\mathbf{S}, \mathbf{x})/\Phi] = 1 \tag{4.4-38}$$

Comparing this with (4.4-3) and (4.4-16), we identify $[\Phi(\mathbf{S}, \mathbf{x})/\Phi]$ with $[F(\omega)/\hbar\omega]$ so that we define,

$$F(\mathbf{S}, \mathbf{x})/\hbar\omega_{\mathbf{x}} = \Phi(\mathbf{S}, \mathbf{x})/\Phi \tag{4.4-39}$$

Although previously, $F(\omega)/\hbar\omega$ was defined as a frequency pdf for spectrum modeling, $F(\mathbf{S}, \mathbf{x})/\hbar\omega_{\mathbf{x}}$ is a more general model suitable for modeling various types of signals and images. Now, comparing Shannon's similarity measure (4.4-37) with Einsteinian likelihood (4.4-4), we identify the two and call it the Shannon–Einsteinian similarity

$$\begin{aligned} L &= \text{pdf}(X|Y) = \Gamma_{\text{W}|XY}/\,\Gamma_{\text{W}|N,\Phi} = \prod_{\mathbf{x}} [F(\mathbf{S}, \mathbf{x})/\hbar\omega_{\mathbf{x}}]^{N_{\mathbf{x}}} \quad \text{or} \\ LL &= \ln\ \text{pdf}(X|Y) = \sum_{\mathbf{x}}\ N_{\mathbf{x}}\ \ln[F(\mathbf{S}, \mathbf{x})/\hbar\omega_{\mathbf{x}}] \end{aligned} \tag{4.4-40}$$

The above equations give the mathematical description of the Einsteinian intuition about the spectrum being a probability density of the frequency. The Einsteinian idea is extended here to a general intensity or power model being a pdf of the image or signal coordinates. The probability density referred to by Einstein is defined in the space of physical states of the world: it is the density of microstates, corresponding to a macrostate of the measured data $\{S(\mathbf{x})\}$, among all possible microstates of the world (compatible with our a priori knowledge about the world). (It is remarkable that Einstein proposed this idea about 20 years before the birth of quantum mechanics!) In addition to formalizing the concept of the Einsteinian likelihood, the above equations relate it to Shannon's concept of mutual information: the logarithm of the Einsteinian likelihood is a measure of mutual information in the model about the data.

4.4.8 Shannon–Einsteinian MFT Dynamics

Shannon–Einsteinian AZ-similarity is defined from (4.4-40) similar to the general AZ-similarity, by using a compositional model for $F(\mathbf{S}, \mathbf{x})$,

$$F(\mathbf{S}, \mathbf{x}) = \sum_{k}\ F(\mathbf{x}|k), \quad k = 1, \ldots, K \tag{4.4-41}$$

where each agent-model $F(\mathbf{x}|k)$ depends on model parameters of kth class, $\mathbf{S}_k$, and the entire vector of all model parameters is a set of all submodel parameters, $\mathbf{S} = \{\mathbf{S}_k\}$. According to (4.4-40), the partial similarity

$$\text{l}(\mathbf{x}|k) = F(\mathbf{x}|k)/\hbar\omega_{\mathbf{x}} \tag{4.4-42}$$

and the pixel similarity

$$\text{l}(\mathbf{x}) = \left[\sum_{k}\ \text{l}(\mathbf{x}|k)\right]^{N\mathbf{x}} \quad \text{or} \quad \text{ll}(\mathbf{x}) = N_{\mathbf{x}}\ \ln\left[\sum_{k}\ \text{l}(\mathbf{x}|k)\right] \tag{4.4-43}$$

Comparing the pixel similarity $\mathrm{ll}(\mathbf{x})$ to the one used previously in Section 4.3, $\mathrm{ll}(n)$, we can interpret $\ln\left[\sum_k \mathrm{l}(\mathbf{x}|k)\right]$ as a single-photon similarity, $\mathrm{ll}(n)$, and $\mathrm{ll}(\mathbf{x})$ being a similarity for pixel $\mathbf{x}$ containing $N_\mathbf{x}$ photons. The Shannon–Einsteinian AZ-similarity between the data set $\{S(\mathbf{x})\}$ and a set of models $\{F(\mathbf{x}|k)\}$ is given by

$$\text{AZ-}LL = \sum_\mathbf{x} N_\mathbf{x} \ln\left[\sum_k \mathrm{l}(\mathbf{x}|k)\right] = \sum_\mathbf{x} N_\mathbf{x} \ln\left[\sum_k F(\mathbf{x}|k)/\hbar\omega_\mathbf{x}\right] \tag{4.4-44}$$

An intuitively appealing interpretation of the above is that the sum over pixels ($\Sigma_\mathbf{x} N_\mathbf{x}$), can be replaced by the sum over individual photons, (Σ_n); then the above similarity "looks" exactly the same as the general AZ-similarity introduced in Section 4.2. And the dynamics maximizing the similarity is just like the general case derived in Section 4.2,

$$\begin{aligned} d\mathbf{S}_k/dt &= \sum_\mathbf{x} N_\mathbf{x} f(k|\mathbf{x})\ \partial \mathrm{ll}(\mathbf{x}|k)/\partial \mathbf{S}_k \\ &= (1/\hbar\omega_\mathbf{x}) \sum_\mathbf{x} N_\mathbf{x}\ f(k|\mathbf{x})\ \partial \ln\ F(\mathbf{x}|k)/\partial \mathbf{S}_k \end{aligned} \tag{4.4-45}$$

$$f(k|\mathbf{x}) = \mathrm{l}(\mathbf{x}|k)/\left[\sum_{k'} \mathrm{l}(\mathbf{x}|k')\right] = F(\mathbf{x}|k)/\left[\sum_{k'} F(\mathbf{x}|k')\right] \tag{4.4-46}$$

with the only difference here being that Σ_n is replaced with $\Sigma_\mathbf{x} N_\mathbf{x}$.

If the Einsteinian Gaussian mixture model is used, as in Section 4.4.3, for the signal-source models,

$$\begin{aligned} F(\mathbf{x}|k) &= \hbar\omega\, A_k\, G(\mathbf{x}|k) \\ G(\mathbf{x}|k) &= (2\pi)^{-D/2}\,(\det\ \mathbf{C}_k)^{-1/2}\ \exp\left\{-0.5\ (\mathbf{x}-\mathbf{x}_k)^{\mathrm{T}}\,\mathbf{C}_k^{-1}\ (\mathbf{x}-\mathbf{x}_k)\right\} \end{aligned} \tag{4.4-47}$$

Again, A_k is a submodel amplitude, $\mathbf{x}_k$ is the submodel mean position vector, and $\mathbf{C}_k$ is the submodel covariance, determining the shape of the submodel in D-dimensional $\mathbf{x}$-space. Term $[A_k\, G(\mathbf{x}|k)]$ is measured in units of photon number, and $G(\mathbf{x}|k)$ is interpreted as the conditional pdf of photons from source k. The estimation equations for the model parameters are

$$\begin{aligned} A_k &= N_k/N, \quad N_k = \sum_\mathbf{x} f(k|\mathbf{x})\ [S(\mathbf{x})/\hbar\omega], \quad N = \sum_\mathbf{x} [S(\mathbf{x})/\hbar\omega] \\ \mathbf{x}_k &= \sum_\mathbf{x} f(k|\mathbf{x})\ [S(\mathbf{x})/\hbar\omega]\ \mathbf{x}/N_k \\ \mathbf{C}_k &= \sum_\mathbf{x} f(k|\mathbf{x})\ [S(\mathbf{x})/\hbar\omega]\ (\mathbf{x}-\mathbf{x}_k)\ (\mathbf{x}-\mathbf{x}_k)^{\mathrm{T}}/N_k \\ f(k|\mathbf{x}) &= F(\mathbf{x}|k)/\left[\sum_\mathbf{x} F(\mathbf{x}|k')\right] \end{aligned} \tag{4.4-48}$$

where $f(k|\mathbf{x})$ has the meaning of the a posteriori Bayes probability that a photon with coordinates $\mathbf{x}$ has originated from the source (or submodel) k. Thus, N_k is the number

of photons from the source k, and N is the total number of photons. As in the previous sections, this set of equations defines a convergent iterative system. Although the Gaussian mixture model described above can model any intensity image, usually, it is practically useful for images composed of relatively few Gaussian "blobs." More complex models that are required for more complicated images are discussed in Chapters 6 and 7.

A model-based neural network architecture implementing the above equations is similar to MLANS (at the top-level); it is considered in Section 4.5. We call it an Einsteinian or Shannon–Einsteinian neural network, or a Maximum Entropy Adaptive Neural System, MEANS.

4.4.9 Historical Roots of Maximum Information and Maximum Entropy Estimation

The information or entropy maximization principle that we used to derive Shannon–Einsteinian similarity is connected to the maximum entropy estimation principle that has being used since the 1950s. There are significant similarities and differences between various approaches and methods under similar names, so it is useful to provide a brief overview of the history and literature concerning the roots of the maximum entropy estimation in statistics. Kullback and Leibler (1951), following Khinchin, developed a measure of information distance that was used to develop Khinchin–Kullback–Leibler estimation approaches, also known as maximum entropy (ME) (Jaynes, 1957), minimum cross-entropy (MCE), and minimum discrimination information (MDI). The ME philosophy was formulated by Jaynes (1957) as "maximally noncommittal with regards to missing information." In ME estimation, the problem is formulated as follows. Estimate a pdf, $q(n)$, given a set of linear constraints on q. Constraints are given in terms of expected values, i.e., of the form

$$\sum_n q(n)\, a(k,n) = 0 \,(\text{or} \geq 0), \quad \text{for } k = 1, \ldots, K \tag{4.4-49}$$

Estimation shall also account for a normalization constraint,

$$\sum_n q(n) = 1 \tag{4.4-50}$$

The ME estimation consists in maximizing the entropy defined as

$$\max\ E; \quad E = -\sum_n q(n)\ \ln\ q(n) \tag{4.4-51}$$

According to the ME philosophy, ME estimates a function $q(n)$ by maximizing its randomness, while satisfying constraints (4.4-49) and (4.4-50). When a prior guess $p(n)$ estimating $q(n)$ is known in addition to the above constraints, MCE is used, which minimizes cross-entropy CE,

$$\min\ CE; \quad CE = \sum_n q(n)\ \ln[q(n)/p(n)] \tag{4.4-52}$$

subject to constraints (4.4-49) and (4.4-50). MCE minimizes "information measure necessary to change $p(n)$ into $q(n)$" subject to the constraints. Note the differences between

Eq. (4.4-52) and our Eqs. (4.4-19) and (4.4-50): in MCE like in ME, the sum is weighted with the sought function $q(n)$, whereas in our definition, the sum is weighted with the measured numbers of photons. Also, a weak point of MCE is that constraints that are used in MCE are of an approximate nature, but they are required to be exactly satisfied. Our approach avoids this undesirable property. But the main difference is in the basic principle: our approach maximizes information in the estimation system about the world.

Shore and Johnson (1980) provided an axiomatic foundation for MCE. They developed requirements for consistent inferences. Their main requirement for consistency is that for statistically independent systems, estimates have to be statistically independent. Classical MCE as well as our definition of the mutual entropy satisfy these conditions (Problem 4.4-7). Shore (1984) has shown that classical ML estimation is equivalent to MCE estimation. He made an important point that the ML estimation is justified from the basic principles only if a model is exactly correct (for some set of parameters), whereas MCE does not rely on this. Thus, he argued that MCE is a more general approach than ML. This is also true for Shannon–Einsteinian similarity described in this chapter.

4.4.10 Likelihood, Information, Ergodicity, and Uncertainty

Information and entropy as well as likelihood are defined in terms of pdfs. The relationship between statistical likelihood on the one hand and information and physical entropy on the other has been addressed throughout this chapter and here we summarize these discussions. First, let us analyze definitions of pdfs considered in statistics, information theory, and physics. In particular, what is the source of uncertainty, pdf of *what* is considered? In statistics, pdf(X) characterizes deviations of the data X from its predictions specified by the model. In information theory and in statistical physics, pdf(X) characterizes the relative frequency of a macrostate X among all macrostates, whereas the model specifies the numbers of microstates. One may wonder if specifying the number of microstates provides an adequate description of a physical system? How does it correspond to considering photon emission and absorption processes? For example, a model for the pdf of pixel intensity could be derived by analyzing the random process of photon emission and absorption in a unit of time within the spherical angle of a pixel. This would require knowledge of the emission and absorption properties of materials, etc.; however, after a relatively long derivation,[9] the procedure will lead to the above equations, which depend only on the numbers of states. In the derivation of the above expressions (4.4-22) for Shannon's similarity (mutual information) we had no need to account for the temporal randomness of the photon emission process. The results are the same, and this is called a principle of ergodicity in statistical physics: in many aspects, temporal randomness and ensemble randomness are equivalent.

The difference between statistical likelihood and Einsteinian likelihood (or equivalently, physical entropy) is not due to considering different physical processes, but to attribution of uncertainty: in statistics—to the unknown variabilities in the data, and in physics—to the unknown microstate of the world.

4.4.11 Forward and Inverse Problems

In statistical physics, a classical paradigm is to find the density of particles by maximizing entropy, *given* the physical model (a *forward* problem). MFT solves an *inverse* problem: given

the observed density of particles (photons), find parameters of the model. The paradigm of an inverse problem is central in statistical estimation, and a natural problem for an intelligent system. A rich body of estimation theory is available for simple inverse problems, when there is just one unstructured source of the data. Complicated inverse problems are characterized by *unknown structure (multiple sources) and unknown parameters* of models. In the past, solutions of complex inverse problems with multiple sources were approached by variations of the previously mentioned Multiple Hypothesis Testing (MHT) algorithm. MHT combines forward modeling with statistical estimation: (1) postulate an association between data and their sources (a hypothesis); (2) estimate parameters of the source model *conditional* on the association hypothesis; (3) solve the *forward* problem; (4) by comparing this solution with the data, estimate an improved association between data and their sources; this is usually accomplished by methods of statistics such as the nearest neighbor; (5) iterate steps 2, 3, and 4 until the solution matches the data. MHT solutions are often prohibitively expensive due to the inherent combinatorial complexity of MHT. Contrary to this, MFT solves an inverse problem without combinatorial complexity by using fuzzy associations between data and their sources. Therefore, statistical estimation performed by Shannon–Einsteinian MFT efficiently solves the physical problem of inverse modeling by combining statistical physics with information theory.

4.5 Modeling Field Theory Neural Architecture

Learning in MFT is determined by a concurrent evolution or adaptation of model parameters $\mathbf{S}_k$ and fuzzy class memberships $f(k|n)$ associating data (input nodes) n with agent-models k. This evolution is given by the dynamic equations for $\mathbf{S}_k$ and $f(k|n)$. Correspondingly, a top level neural architecture consists of two subsystems, an association subsystem and a modeling subsystem (Fig. 4.5-1). An association subsystem implements $f(k|n)$ dynamics given either directly, according to its definition equations [(4.1-19), (4.2-4), (4.3-13), (4.4-8), (4.4-46), and (4.4-48)], or according to the corresponding dynamic equations [(4.2-5), (4.3-26), or (4.3-32)]. Modeling subsystem implements $\mathbf{S}_k$ dynamics given by any of the following equations [(4.2-3), (4.2-7), (4.3-12), (4.3-19), (4.3-21), (4.3-24), (4.4-7), (4.4-13), (4.4-14), (4.4-15), (4.4-45), and (4.4-48)].

The following chapters consider a number of specific MFT models and variations of the above equations, implemented for various applications. General characteristics of these implementations follow from the general MFT equations derived in this chapter and are summarized here. In a digital implementation on a general purpose computer, the continuous dynamic equations are implemented in finite steps, according to the following algorithm:

1. at $t = 0$, initialize parameter values $\{\mathbf{S}_k(t = 0)\}$ according to available a priori knowledge of the problem;
2. using these parameter values, compute models $\mathbf{M}_k(\mathbf{S}_k)$;
3. compute partial similarities $l(n|k)$ and fuzzy memberships $f(k|n)$;
4. compute derivatives of models and partial similarities;
5. compute $[d\mathbf{S}_k/dt]$ according to the corresponding dynamic equation; then compute parameter values at the next iteration time $t + dt$, according to

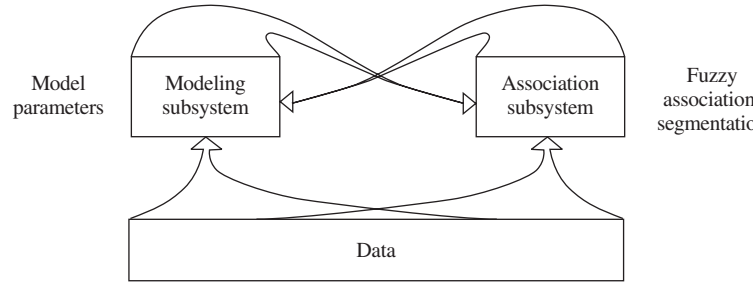


Figure 4.5-1 Modeling field theory neural network architecture; top level. An association subsystem computes weights that associate data with models. Modeling subsystem estimate parameters of the models.

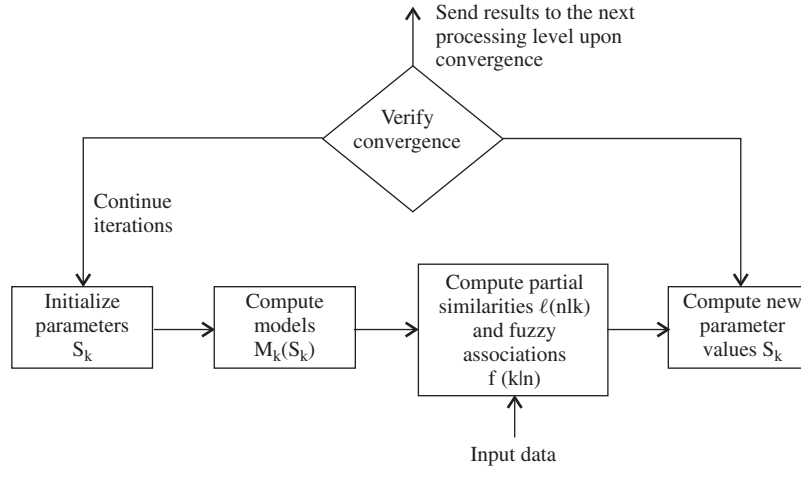

Send results to the next processing level upon convergence
Verify convergence
Continue iterations
Initialize parameters S_k
Compute models $M_k(S_k)$
Compute partial similarities ℓ(nlk) and fuzzy associations f (kln)
Compute new parameter values S_k
Input data

Figure 4.5-2 MFT algorithm.

$$\mathbf{S}_k(t+dt) = \mathbf{S}_k(t) + [d\mathbf{S}_k/dt]$$

6. at $t > 0$, verify convergence according to a predetermined criterion, e.g., compute the log-similarity, *LL*, and check if its change is below threshold

$$LL(t) - LL(t-dt) < \text{threshold}$$

if the convergence criterion is not satisfied, continue iterative estimation: go to step 2; if the convergence criterion is satisfied, stop iterative estimation: go to the next level of decision making.

This iterative algorithm is illustrated in Fig. 4.5-2. Figures 4.5-1 and 4.5-2 show alternative views on the MFT operations. The mapping between these two views is straight-

forward: steps 1 and 6 are implemented by separate initiation and termination subsystems, not shown in Fig. 4.5-1; steps 2, 4, and 5 are implemented in the modeling subsystem; step 3 is implemented in the association subsystem. For complicated models involving spatiotemporal model propagation, separate subsystems might be used for steps 2 and 4.

The neural architecture and learning dynamics described here are related to the MFT models in an essential way: they are based on these models. For this reason, the architecture of Fig. 4.5-1 is called Model-Based Neural Network (MBNN). The terms MFT and MBNN are used interchangeably, whereas the terms MLANS, MEANS, ENN, etc. are reserved for specific implementations using likelihood or information similarities.

4.6 Convergence

4.6.1 Aspects of Convergence

MFT is a convergent dynamic system whose stationary points correspond to the maximal values of AZ-similarity. This means that in finite time the MFT model parameters come arbitrarily close to the stationary point ($dS/dt = 0$), and AZ-L similarity comes arbitrarily close to its maximum. To establish convergence, a convergence criterion should be defined. This can be done by monitoring the rate of change of the similarity measure: when this rate falls below a threshold, a convergence is declared. Sometimes monitoring model-parameter rate of change is more convenient, because these are physical quantities, and thresholds for their rates of change could be more easily determined for specific applications. Each parameter rate of change could be monitored separately; alternatively, a single function of model parameters could be defined determining an average convergence. These various techniques will be illustrated in the following chapters.

In general, MFT dynamics guarantees attaining a local maximal value. AZ-L is a complex nonlinear function with a large number of maxima. The main way MFT looks for the global maximum is selecting initial parameters in such a way that the MFT models $\mathbf{M}_k$ are very fuzzy initially and their degree of fuzziness is reduced during convergence to an appropriate degree. This is achieved by selecting large initial values of covariances. In many cases this is sufficient for finding either the global maximum or a sufficiently good solution. When it is not sufficient, it can be supplemented by running the MFT system multiple times with different initial states (parameter values) and then selecting the result with the highest similarity.

A more general way of finding the global maximum (or a sufficiently good one) is to consider a single-level MFT system as a part of a multilevel hierarchical system. A higher level interprets convergence results of a lower level. This interpretation accounts for the nature and meaning of the MFT models: each of the kth class models corresponds to an object or process recognized within the broader context of multiple models continually adapting to the continuous stream of data. So the problem of global vs. local maxima is related to finding objects present in the data without errors, and, therefore, is related to the problem of the true number of objects (or classes). When the number of object-classes

varies in the estimation process, the AZ-similarities defined in previous sections should be modified as discussed in Section 4.7.

Continuous operation alters the definition of convergence. We should account for two time scales: an internal time scale determined by the parameter dt in MFT equations and an external, real time. When internal MFT evolution is much faster than real-time changes, the conventional convergence criteria are useful, but when the two time scales are commensurate, only local measures of convergence could be defined. The higher level makes decisions concerning validity of correspondence between the data and a set of active class-models that is being estimated at a given moment in real time. A decision could be that there is no correspondence between a particular model and the data, and this model should be eliminated from the processing stream, or that a particular model matches the data well so that conclusions should be drawn from the presence of this particular object, or that another model should be activated for modeling a particular piece of data. All of these cases are considered in the following chapters.

4.6.2 Proof of Convergence

We will discuss here two proofs, first related to the gradient ascent and second, applicable to the system of parallel dynamic equations for parameters and neural weights. The gradient ascent leads to a continuous monotone increase of the similarity. This can be shown as follows. Consider change in similarity LL during time dt due to adaptation of a set of parameters $\{\mathbf{S}_k\}$:

$$dLL = \sum_k [\partial LL/\partial \mathbf{S}_k]\ [d\mathbf{S}_k/dt]\, dt \tag{4.6-1}$$

where

$$d\mathbf{S}_k/dt = \partial LL/\partial \mathbf{S}_k \tag{4.6-2}$$

Substituting the second into the first, we obtain,

$$dLL/dt = \sum_k [\partial LL/\partial \mathbf{S}_k]^2 \geq 0 \tag{4.6-3}$$

If the similarity is finite for all values of $\mathbf{S}_k$, then, within a finite number of steps $d\mathbf{S}_k$, the MFT system will come to a point where dLL increases are arbitrarily small. Such a point is either a maximum or a point of inflection, and $d\mathbf{S}_k$ approaches zero near such a point, thus it is a stationary point. It remains to analyze conditions under which a similarity is finite. It is sufficient to consider partial pixel similarities, $l(n|k)$. For parametric models considered in this book, $l(n|k)$ could become infinite only if its covariance goes to zero ($\det \mathbf{C} \to 0$). This is equivalent to $l(n|k)$, for object-class k, concentrating on a single data point n, or $l(n|k)$ being nonzero in the infinitely small vicinity of the data point n. These infinities of similarity can be prevented by forcing covariances to be larger than some small predetermined amount. Thus, the similarity is finite and convergence is always attained in a finite number of steps. This proof is applicable to a "straightforward" gradient ascent given by any of the dynamic

equation for the model parameters [Eqs. (4.2-3), (4.2-7), (4.3-12), (4.3-19), (4.3-21), (4.3-24), or (4.4-7), (4.4-45)], while the neural weights $f(k|n)$ are computed according to either equations [(4.1-19), (4.2-4), (4.3-13), or (4.4-8), (4.4-46), (4.4-48)].

Consider now a convergence proof suitable for the pair of dynamic equations determining the evolution of both parameters and neuronal weights [Eqs. (4.2-3), (4.2-4), (4.2-7), (4.3-21), (4.3-22), (4.3-24), or (4.4-48)]. As in the pervious proof, it is sufficient to show that the MFT dynamics leads to increasing (nondecreasing) similarity. We modify the notations by adding a parameter t describing the dynamic evolution of corresponding quantities in the process of convergence, e.g., $\mathrm{l}(n|k,t)$ is a partial similarity $\mathrm{l}(n|k)$, computed with the parameter values $\mathbf{S}_k(t)$, estimated at time t, and $\mathrm{ll}(n|k,t) = \ln \mathrm{l}(n|k,t)$, etc. First, let us prove the following two lemmas.

LEMMA 4.6.1. Given $\Sigma_k\ q_k = 1$, $\Sigma_k\ p_k = 1$. The $\min_{p_k}\ [\Sigma_k\ q_k\ \ln(q_k/p_k)] = 0$, and it is attained at $p_k = q_k$.

Proof. Using the method of Lagrange multipliers, the minimum, given the constraint, is found by

$$\partial/\partial p_k \left[\sum_k q_k \ln(q_k/p_k) + \lambda\,(p_k - 1) \right] = 0 \tag{4.6-4}$$

This leads to $-q_k/p_k + \lambda = 0$, or $q_k = \lambda p_k$. Summing up this expression over $k = 1, \ldots, K$, and using the constraint, we get $\lambda = 1$, which proves the lemma.

LEMMA 4.6.2. Dynamic equations [(4.2-5),(4.3-20), and (4.3-22)] for the neural weights (fuzzy class memberships) preserve the definition

$$f(k|n,t) = \mathrm{l}(n|k,t)/[l(n|1,t) + \cdots + \mathrm{l}(n|K,t)] \equiv \mathrm{l}(n|k,t)/\mathrm{l}(n|t) \tag{4.6-5}$$

at all times t.

Proof. This follows from the dynamic equation for $f(k|n,t)$ being the time derivative of the right-hand side of Eq. (4.6-5) and the initial condition given by this equation at $t = 0$. It also follows that

$$\sum_k f(k|n,t) = 1 \tag{4.6-6}$$

at all times, t.

THEOREM 4.6.3. A system of dynamic equations for $d\mathbf{S}_k$/dt and $df(k|n)/dt$ of the type (4.2-3), (4.2-4), (4.3-21), (4.3-22), (4.3-24), or (4.4-48) is a converging system; it converges to a stationary point of log-similarity AZ-*LL* (that is, to a point, where dAZ-*LL*/$d\mathbf{S}_k = 0$).

Proof. Examine a change of log-similarity from t to $t + dt$:

$$
\begin{aligned}
\text{AZ-}LL(t+dt) - \text{AZ-}LL(t) &= \sum_{n=1}^{N} [\text{ll}(n|t+dt) - \text{ll}(n|t)] \\
&= \sum_{n=1}^{N} \left\{ \sum_{k=1}^{K} f(k|n,t) \right\} [\text{ll}(n|t+dt) - \text{ll}(n|t)] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} f(k|n,t)\ [\text{ll}(n|k,t+dt) \\
&- \ln\ f(k|n,t+dt) - \text{ll}(n|k,t) + \ln\ f(k|n,t)] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} f(k|n,t)[\ln\ f(k|n,t) - \ln\ f(k|n,t+dt)] \\
&+ \sum_{n=1}^{N} \sum_{k=1}^{K} f(k|n,t)[\text{ll}(n|k,t+dt) - \text{ll}(n|k,t)]
\end{aligned}
\tag{4.6-7}
$$

In the first line here we used an identity Eq. (4.6-6); in the second line we used the logarithm of Eq. (4.6-5). The third line is nonnegative because of Lemma 4.6.1 (its min value is 0). The last line is also nonnegative (see Problem 4.6-1). Thus, *dLL/dt* is nonnegative. This completes the proof of the MFT convergence at least to a local maximum (or inflection point) of the similarity.

A modification of the expectation-maximization (EM) algorithm used in several problems in this chapter to obtain the estimation equations can be formulated as follows. An iterative maximization of AZ-similarity can be achieved by maximizing at every iteration, $t + dt$, the following expression:

$$
\max_{\mathbf{S}_k(t+dt)} \left\{ \sum_n f\ [k|n, \mathbf{S}(t)]\ \ \text{ll}[n|k, \mathbf{S}(t+dt)] \right\} \tag{4.6-8}
$$

Here, fuzzy memberships $f(k|n, \mathbf{S}(t))$ are evaluated using the parameter values $\mathbf{S}(t)$ estimated at the previous iteration, t.

COROLLARY 4.6.4. The EM procedure converges to a maximum of log-similarity AZ-*LL*.

Proof. The EM procedure ensures that the last line in (4.6-7) is nonnegative, and since the previous line is nonnegative, then *dLL* is nonnegative as well.

4.7 Learning of Structures, AIC, and SLT

Learning structural characteristics of the internal model often is considered a different problem from parameter adaptation. This is not so in MFT: a structure of the internal model

is determined by the number and interaction of submodels (which we also call object-models, components, or agents). Therefore, learning of a structure is an integral part of the MFT learning. Estimating the number of agents or submodels from the data requires modifications of the procedures discussed in previous sections. These modifications are discussed here.

There are two basic procedures for estimating the number of agents, depending on the role of agents and their fidelity. First, each agent submodel can be used to model a distinct object or physical source of signals. This is appropriate when the submodel fidelity is adequate for the purpose. Second, several agents can be used for a single object or signal source. This is appropriate when the a priori knowledge is insufficient for accurate modeling, and several submodels are used to account for expected variabilities. In the first case, the actual presence or absence of the object can be decided by computing an agent activation, $a(k) = \sum_n f(k|n)$, and comparing it against a decision threshold. The agent-concept activation is the degree of recognition of the object or concept by the MFT system (see a related discussion in Section 4.2.4). In the second case, the decision is more subtle: how many sub-models should be used for accurate modeling? And what does constitute an accurate modeling? The difficulty is related to the fact that with a sufficiently large number of parameters, any data can be fit very closely, so that the partial similarities and activations can become very large, even if the modeled objects are not present. Therefore, it is necessary to balance maximization of similarity against the number of parameters in the model.

Mathematically, the above problem can be formulated as follows. Let us consider the maximal value of likelihood attained in the process of parameter estimation (which maximizes the similarity). This estimated maximal likelihood value is a random quantity, depending on the available data. How well does this estimated value represent the true maximal value of the likelihood? The maximum likelihood estimation is *asymptotically unbiased*, that is, for a sufficiently large amount of data, the estimated value, on average, is close to the true value. The "sufficiently large amount of data" supposes a large amount of data compared to the number of parameters. If the number of parameters varies it might grow indefinitely, so that no amount of data is "sufficiently large." Thus, the *asymptotically unbiased* property of the ML has to be reexamined. It turns out that the estimated maximal loglikelihood value is biased upward: its expectation $E\{LL\}$ is above the true value, LL_0, and this bias is proportional to the number of parameters, N_{par},

$$E\{LL\} = LL_0 + N_{\text{par}}/2 \tag{4.7-1}$$

Therefore, if the number of parameters varies in the estimation process, the proper quantity to be maximized is not the log likelihood, but the log likelihood corrected for the bias. This quantity is called the Akaike Information Criterion, AIC,

$$AIC = LL - N_{\text{par}}/2 \tag{4.7-2}$$

This correction is equally applicable to Bayesian and Einsteinian likelihoods. (And it is related to a well-known fact in statistical physics that every degree of freedom, on average, consumes $kT/2$ energy, where k is the Boltzmann constant and T is temperature; the AIC states that every parameter that is a degree of freedom in an estimation system has to be attributed one-half of the log-likelihood value.) It follows, that if the number of

parameters varies in the estimation process, similarity measures introduced in this chapter should be corrected,

$$\text{AZ-}LL \rightarrow \text{AZ-}LL - N_{\text{par}}/2 \tag{4.7-3}$$

Throughout the book we will assume this correction, when appropriate.

AIC criterion "penalizes" likelihood proportionately to the number of adaptive model parameters. The penalty is relative to the number of observations or data vectors, N, used for parameter estimation (training), because *AZ-LL* is proportional to N. This procedure is appropriate for a relatively large number of training observations ($N \gg N_{\text{par}}$), because it is based on an *asymptotic* estimation of the likelihood bias. For learning from small samples ($N > N_{\text{par}}$, but not "$\gg$"), it has been empirically observed that AIC does not sufficiently "penalize" the likelihood function. In simple cases, such as linear classification (and linear regression), it is possible to derive a more accurate "penalty" procedures for learning from small samples. This is a subject of statistical learning theory, SLT (Section 2.7). The MFT learning developed in this book is based on utilizing complex a priori models with small number of parameters ($N_{\text{par}} \ll N$). This usually results in complicated nonlinear problems, so that SLT penalization procedures are not applicable. Combining SLT and MFT is a challenging problem for future research.

4.8 Instinct of World Modeling: Knowledge Instinct

It is an intriguing possibility to consider a biological interpretation of the MFT dynamics, maximizing the similarity between the internal model and the outer world data. Mathematical, psychological, and philosophical considerations suggest that the mind does not simply react to sensory data by acting out in the outer world. Complex internal representations (models) are utilized in order to *understand* the world: to recognize objects, their relationships, and possible interactions with them. Knowledge is represented by internal models. Acting out in the world is often very indirectly related to sensory stimuli, and the internal model often takes on a "life" of its own. Knowledge, or a good correspondence between the model and the world, is so important that there have to be very basic biological mechanisms driving toward regular or even constant improvements of this model. For example, dolphins placed in a new environment first map it out acoustically with specific sound signals; then they can easily find a new object placed in the water. Many other forms of exploratory behavior of animals can be explained by assuming a basic instinct or drive to improve the internal model. The MFT similarity maximization mechanism is a possible mathematical description of that instinct or drive to improve the world model, the instinct for knowledge. This discussion is continued in Chapter 10, where the functioning of the mind is analyzed on a system level, combining mathematical and philosophical inquiry.

4.9 Summary

This chapter described the modeling field theory (MFT), a dynamic learning system, whose adaptation or learning is based on a priori models as well as on empirical data. The MFT

combines internal models with similarity measures utilizing available knowledge with various degrees and types of uncertainty and variability: deterministic, stochastic, and fuzzy. In MFT, concepts or classes emerge in the process of learning; the initial state of MFT models can be characterized as fuzzy concepts; during MFT learning, the degree of fuzziness is reduced and crisp (or less fuzzy) concepts emerge. MFT is an organization of intelligent agents. Every agent has its own internal model and the MFT system maximizes the similarity between all the models and the input data. In the process of similarity maximization, agents compete for the evidence in the data, which supports their models, while adapting their models to the data. The learning process therefore includes segmentation or association of data with various agent-models. Models are the internal representations of structures or patterns in the data and are the mathematical descriptions of concepts of mind. Similarity measures are the evaluative signals and are the mathematical descriptions of emotions; these emotions correspond to satisfaction of the instinct for knowledge.

Utilization of a priori models solves the problem of learning from a limited amount of data. Models for particular applications are to be designed to strike a balance between the available a priori knowledge and available training data. The mathematical apparatus of MFT combining fuzzy logic with adaptivity and apriority resolves the long-standing issue of seemingly inescapable combinatorial complexity of the past approaches to modeling the intellect, discussed in Chapter 2. The concept of the model-based neural network corresponds to the Aristotelian theory of mind and fulfills the McCulloch vision of learning based on complex a priori neural structures.

We discussed a correspondence between MFT and the Aristotelian theory of mind (theory of Forms). According to Aristotle, the a priori contents of mind are Forms, which contain concepts as potentialities, and which become concepts in the process of learning when meeting the matter. MFT describes Forms mathematically as fuzzy agent-concepts, which emerge as crisp concepts in the process of learning through interaction with empirical data. In his theory of Forms, Aristotle criticized Plato's theory of mind based on eternally true, ready-made Ideas (crisp concepts). But we discussed that the Aristotelian theory of *logic* was more suitable for eternal Ideas of Plato than for fluid adaptive Aristotelian Forms. Reliance on Aristotelian logic caused difficulties in past mathematical attempts to model the intellect. This difficulty is resolved in MFT by using adaptive model-based fuzzy logic. Thus, 2300 years since Aristotle, the logical foundation for the Aristotelian theory of mind is provided by the fuzzy logic originated by Zadeh.

The dynamics of MFT, determining the learning equations, is obtained by maximizing the AZ-*L* similarity between the input data and a set of models. This similarity measure includes all segmentations/associations. Thus, MFT accomplishes concurrently the model estimation and segmentation (concept formation). We considered two types of similarities based on maximization of likelihood (ML) and mutual information (MMI). The ML is a fundamental statistical principle. In addition to the intuitively appealing notion of "the most likely," it has advantages of being asymptotically unbiased and efficient. Often, the ML estimation leads to the fastest possible learning. Nonasymptotically, when the amount of data is small, or when the environment changes so fast that there are not enough data to attain asymptotic accuracy, the ML principle is not guaranteed to be "the best." For some problems, in a nonasymptotic region, there are better estimation techniques than the ML. For example, in Chapter 6 Einsteinian likelihood results in a better spectrum estimation than the ML, even with exact models. However, such techniques

are "rarities"; there is no *general* estimation technique better than the ML *when models are accurate.*

The model accuracy requirement could be an important limitation to the ML estimation. This requirement is not a trivial one: for example, it is often possible to specify a very flexible model with a very large number of parameters, so that it can fit the data perfectly. Nonparametric techniques, such as feedforward or nearest neighbor neural networks, can be used for this approach. Perfect fitting of available data by using unrestrictive models with an excessive number of parameters often leads to no valid generalization: results of learning are not applicable to new data. The model accuracy requirement for the ML estimation is a requirement of the *true* model: the model should be the true representation of the reality. This is rarely achievable in practice: our models are always only approximations to the real world. Note a subtlety here: the ML allows for and assumes the deviations of data from models, but these deviations should be random. Quite often deviations of data from models can be considered random, even if not really so. Therefore, the ML estimation often is quite appropriate even for approximate models.

Another estimation principle introduced in this chapter is based on the maximum mutual information (MMI) between the model and the data. When the model is an accurate physical model of the world, the mutual information can be interpreted as a likelihood. This is a different type of likelihood than is usually considered in statistics, and we call it the Einsteinian likelihood. Although classical statistical likelihood attributes uncertainty to the unknown causes of deviations between the model and the data, Einsteinian likelihood attributes uncertainty to the unknown microstate of the world. The MMI principle, equivalent to Einsteinian likelihood, is founded on extracting maximum information from the data based on the available a priori knowledge (good or bad) captured in the adaptive a priori models, *even if models are approximate.* Therefore, MMI can be better suitable for approximate models than the ML.

Finally, we touched on a possible biological interpretation of the MFT internal dynamics based on similarity maximization, as an instinct or internal drive for learning.

The following chapters consider further development and applications of the two versions of the general MFT theory. Chapter 5 considers Maximum Likelihood Adaptive Neural System (MLANS) utilizing the Bayesian likelihood similarity and the maximum likelihood estimation principle. Chapter 6 considers the Shannon–Einsteinian Neural Network (ENN), utilizing Shannon's information similarity and the maximum mutual information estimation principle.

NOTES

1. Examples of deformable models of continuous geometric patterns were discussed in Terzopoulos et al., (1988). It is a straightforward exercise to rewrite formally the above equations for the continuous index k, by substituting all sums with integrals, similar to what we did in Eqs. (4.2-9) and (4.2-10). It is also possible to define appropriate models and measures of similarity so that the infinite number of models and parameters can be learned from finite training data (Perlovsky, 1996b; Perlovsky et al., 1997b).
2. Central limit theorem states that under a wide range of conditions, a sum of random variable has a Gaussian pdf. This theorem establishes a dominant position of the Gaussian distribution

in probability and statistics. The conditions of this theorem do not hold if several *deterministic* processes are present, e.g., if measured values are random deviations from two deterministically different values.

3. A pdf normalization, $\int \text{pdf}(\mathbf{x}_n|H_k)d\mathbf{x}_n = 1$, usually is accounted for in the definition of the parametric shape of the pdf. It is an identity, satisfied for any parameter values, and we do not need to account for this constraint with Lagrange multipliers. A similar procedure can be used for $P(k)$, e.g., by defining $P(k) = P'_k/(\Sigma_{k'} P'_{k'})$, and treating P'_k as independent parameters.
4. Direct interaction among photons is very weak and is not perceptible except by using specialized equipment designed for this purpose. Therefore, photons produced by most objects or systems reflecting or emitting light are not correlated. In lasers, photon states are correlated due to interaction between photons and electronic states. But even in lasers, after being emitted, photons do not correlate. Most of the observed dependencies among properties of photons are due to deterministic relationships, which we intend to model with our $F(\omega)$ models.
5. Again, as in Section 4.3 on MLANS, fuzziness of the models is controlled by the variance, σ_k^2. Initial fuzziness has to correspond to uncertainty in ω_k. This can be achieved by modifying (4.4-15) as follows: $\sigma_k^2 = \sigma 0_k^2 \exp(-ct) + \sigma 1_k^2$, where $\sigma 0_k^2$ is the initial large variance, $\sigma 1_k^2$is the ML variance estimation given by (4.4-15), t is the iteration number (internal MFT time), and c is a parameter controlling the convergence of fuzziness.
6. The following computation is performed for a "classical" limit, when $\Phi_\omega \gg N_\omega$. Therefore, it may be inaccurate for lasers, still, it would be applicable to most of the images formed using scattered laser light. Also, considering photons with a given frequency as indistinguishable may not be correct if their polarization is also measured. But, as seen from the following, the denominator $(N_\omega!)$ does not affect the MFT estimation equations. So, the indistinguishability considerations do not affect the estimation equations.
7. This needs clarification for those familiar with Shannon's definition of information $I = H(x) H(y) - H(x, y)$, where entropy $H(x) = -\ln P(x)$, and $P(x)$ is the probability of x. The probability $P(x)$ can be interpreted as follows. Denote $\Gamma(X|x)$ the number of states (or equiprobable outcomes) in system X compatible with data x (given data x), and $\Gamma(X|o)$ the total number of states in system X (given no information on X). Then, $P(x) = \Gamma(X|x)/\Gamma(X|o)$, $P(x, y) = P(x|y)P(y) = \Gamma(X|x, y)\Gamma(Y|y)/[\Gamma(X|o)\Gamma(Y|o)]$, $H(x) = \ln[\Gamma(X|o)/\Gamma(X|x)]$, and $I = \ln[\Gamma(X|o)\Gamma(Y|o)\Gamma(X|x, y)\Gamma(Y|y)/(\Gamma(X|x)\Gamma(Y|y)\Gamma(X|o)\Gamma(Y|o)] = \ln[\Gamma(X|x, y)/(\Gamma(X|x)]$. Thus, Shannon's information I is the information in data y on system X given data x. Comparing the last expressions for entropy and information, we conclude that entropy $H(x)$ is the negative information in data x on system X. The following sections elaborate on this line of argument in detail.
8. When counting states in quantum mechanics, indistinguishable states that are characterized by the same set of quantum numbers are treated differently from distinguishable states. If an index $\mathbf{x}$ refers to a complete set of the photon quantum numbers, so that photons within each "pixel" $\mathbf{x}$ are indistinguishable, the number of photon states is reduced by the number of permutations of indistinguishable photons.
9. An approach could be to model the photon emission as a random process, leading to Poisson density, $\text{pdf}(N_\omega) = \lambda \exp(-\lambda N_\omega)$, for the number of photons emitted in a unit time. Note that this expression is a falling function of N_ω and, thus, it is qualitatively different from Eq. (4.4-9). To properly account for the dependence on the number of photons, the above Poisson density has to be multiplied by the phase space associated with N_ω photons, which is proportional to a N_ω-dimensional integral over individual-photon degrees of freedom. The resulting density, in case of classical systems with many photons, will have an extremely sharp peak at the mean value of N_ω. Because of the sharpness of this peak, it is sufficient to consider only the mean value N_ω. Therefore, a standard statistical approach attributing randomness to N_ω values cannot be justified based on a Poisson randomness of the physical process of photon emission. Therefore, any uncertainty in

N_ω value is not due to a Poisson randomness, but to uncertainty about the physical properties of objects and numbers of the photon states. Our approach based on Einsteinian likelihood models the distribution of N_ω (as a function of ω), and is related to varying λ and the number of states as a function of frequency.

BIBLIOGRAPHICAL NOTES

This section contains additional bibliographical information that was not explicitly referenced in the text.

Bayes theory (Bayes, 1763; Press, 1989; Jaynes, 1994).
Gradient methods (Press et al., 1989).
Maximum likelihood estimation (Cramer, 1946).
Development and applications of similarity measures, MFT, MBNN, MLANS, MEANS, ENN (Perlovsky, 1987b, 1988b 1994b, 1995, 1996b, 1997d; Perlovsky and Plum, 1991; Perlovsky et al., 1994, 1995a,b, 1996, 1997a,b,c).
Other types of neural networks incorporating statistical mixture models similar to MLANS: SPNN (Streit and Luginbuhl, 1990, 1994); HME (Jacobs et al., 1991; Jordan and Jacobs, 1994); HMD and POEM (Kumar and Manolakos, 1996; Baggenstoss, 1997). Probabilistic neural network, PNN (Specht, 1990) estimates pdf using a nearest neighbor type Parzen estimation.
Akaike information criterion (Akaike, 1973).
Estimation-maximization algorithm (Dempster et al., 1977).
Statistical physics, counting numbers of states and entropy calculation (Landau and Lifshitz, 1980; Sakurai, 1985).
Dolphin behavior (Greineder, 1995).

PROBLEMS

Section 4.1

4.1–1 Compare the description of the levels of model representation to the description of the intelligent tracker in Section 1.1.4.

Section 4.2

4.2–1 Derive Eq. (4.2-5). *Hints*: Use (4.1-17); show that $\delta_{kk'}$ comes from the derivative of the numerator in (4.2-4) and the rest comes from the denominator.

4.2–2 Evaluate the derivative $\partial \ln\left[\text{pdf}\,(\mathbf{x}_n|H_k)\right]/\partial \mathbf{M}_k$ for Gaussian pdf, Eq. (4.3-6). Show that $\partial \ln\left[\text{pdf}\,(\mathbf{x}_n|H_k)\right]/\partial \mathbf{M}_k = \mathbf{D}_{nk}^{\mathrm{T}}\,\mathbf{C}_k^{-1}$, where $\mathbf{D}_{nk} = \mathbf{x}_n - \mathbf{M}_k$. Use symmetry of the matrix $\mathbf{C}_k^{-1}$.

Section 4.3

4.3–1 Let x_n and x_{n-1} be correlated with a correlation coefficient r. Consider model $M_k = E\{x_n|x_{n-1}\}$, that is, given by the linear regression of x_n on x_{n-1}, Eq. (1.4-21). Show that $(x_n - M_k)$ is uncorrelated with x_{n-1}. *Hints*:

1. According to (1.4-14), $M_k = \bar{x}_n + (x_{n-1} - \bar{x}_{n-1})\, r\, \sigma_n/\sigma_{n-1}$, where σ_n and σ_{n-1} are standard deviations.
2. Compute the correlation between x_{n-1} and $(x_n - M_k)$: $E\{(x_{n-1} - \bar{x}_{n-1})\,(x_n - M_k)\}$ by opening parentheses inside { }.
3. Recollect definitions of standard deviations and correlations from Chapter 1, and show that all items in $E\{\ \}$ cancel each other.

4.3–2 Prove that class rates satisfy the constraint (4.3-8). *Hint*: Take $\int d\mathbf{x}_n$ from each side of (4.3-7). Recollect the pdf normalization from Chapter 1: $\int d\mathbf{x}_n\ \text{pdf}\,(\mathbf{x}_n) = 1$ and $\int d\mathbf{x}_n\,\text{pdf}\,(\mathbf{x}_n|H_k) = 1$.

4.3–3 Derive the rate estimation Eq. (4.3-14). *Hints*:

1. Follow the derivation of (4.3-12), compute $\partial \text{AZ-}LL'/\partial P(k)$ and obtain

$$dP(k)/dt = \partial \text{AZ-}LL'/\partial P(k) = \sum_n f(k|n)\,\partial \ln P(k)/\partial P(k) + \lambda$$
$$= \sum_n f(k|n)/P(k) + \lambda$$

2. At the maximum, $\partial \text{AZ-}LL'/\partial P(k) = 0$ and $dP(k)/dt = 0$, therefore,

$$\sum_n f(k|n)/P(k) + \lambda = 0 \quad \text{or} \quad P(k) = -\sum_n f(k|n)/\lambda$$

3. Take a sum of this over k, use the constraint (4.3-8), use $\Sigma_k\ f(k|n) = 1$, and denote the total number of pixels $N = \Sigma_n\ 1$. Obtain,

$$1 = -N/\lambda \quad \text{or} \quad \lambda = -N, \quad \text{and} \quad P(k) = \sum_n f(k|n)/N$$

Section 4.6 contains a proof that using this equation in place of the gradient ascent also leads to a convergent procedure.

4.3–4 Rewrite Eq. (4.3-21) for a case of $C_{ijk} = c1 \cdot \delta_{ij}$ and $M_{ik}^{:a} = c2 \cdot \delta_{ai}$.

4.3–5 Analyze the above result for $c2 = 1$, and $f(k|n) = (0 \text{ or } 1)$ for some k. Show that $\mathbf{S}_k = \mathbf{M}_k$ leads to $M_{ik}^{:a} = \delta_{ai}$. Show that the stationary point ($d\mathbf{S}_k/dt = 0$) corresponds to the standard estimation for $\mathbf{M}_k$ as an average of those $\mathbf{X}_n$ for which $f(k|n) = 1$.

4.3–6 Derive the modification of the Newton gradient equations as follows. Find an extremum of $f(s)$ by taking at each iteration an extremum of the second-order Taylor expansion of this function: $f(s) = f_0 + (s - s_0)\,f_0' + 0.5(s - s_0)^2 f_0''(s - s_0)$.

4.3–7 Verify that the two Eqs. (4.3-21) and (4.3-24) determine incremental changes of the vector $\mathbf{S}_k$ in an approximately same direction. [*Hint*: using positive definiteness of $\mathbf{C}^{-1}$, show that the dot-product $[(d\mathbf{S}_k/dt)_{4.3-21} \cdot (d\mathbf{S}_k/dt)_{4.3-24}] > 0$; a matrix $\mathbf{C}$ is called positive definite if $\mathbf{x}^{\mathrm{T}}\mathbf{C}\mathbf{x} > 0$ for any $\mathbf{x}$.]

4.3–8 Verify that in Eq. (4.3-24), dt is dimensionless.

4.3–9 Show that the second item in Eq. (4.3-25) gives the ML estimation of the covariance matrix. *Hints*.

1. Derive estimation equations for $\mathbf{C}_k^{-1}$, because they are somewhat simpler than for $\mathbf{C}_k$. The ML estimation is given by $\partial \text{AZ-}LL'/\partial \mathbf{C}_k^{-1} = 0$.

2. Compute $\partial \text{AZ-}LL'/\partial \mathbf{C}_k^{-1}$ [by following the derivation of (4.3-12)]. Additional complications are caused by the constraints that the covariance matrixes are symmetrical. But, first, let us ignore this fact. Use the identities (*which are not correct for symmetrical matrixes*) $\partial \left(\ln \det \mathbf{C}_k^{-1}\right)/\partial \mathbf{C}_k^{-1} = \mathbf{C}_k$, $\partial \left(\mathbf{D}_{nk}^{\mathrm{T}} \mathbf{C}_k^{-1} \mathbf{D}_{nk}\right)/\partial \mathbf{C}_k^{-1} = \mathbf{D}_{nk}^{\mathrm{T}} \mathbf{D}_{nk}$, and obtain

$$\partial \text{AZ-}LL'/\partial \mathbf{C}_k^{-1} = \sum_n f(k|n) \left[0.5\mathbf{C}_k - 0.5\mathbf{D}_{nk}^{\mathrm{T}} \mathbf{D}_{nk}\right] = 0 \tag{P1}$$

   where $\mathbf{D}_{nk} = \mathbf{X}_n - \mathbf{M}_k$. This gives the second item in Eq. (4.3-25).

3. Repeat the above using the correct equations accounting for the symmetry of $\mathbf{C}$:

$$\begin{aligned} \partial \left(\ln \det \mathbf{C}_k^{-1}\right)/\partial \mathbf{C}_k^{-1} &= 2\mathbf{C}_k - \text{diag}\mathbf{C}_k, \partial \left(\mathbf{D}_{nk}^{\mathrm{T}} \mathbf{C}_k^{-1} \mathbf{D}_{nk}\right)/\partial \mathbf{C}_k^{-1} \\ &= 2\mathbf{D}_{nk}^{\mathrm{T}} \mathbf{D}_{nk} - \text{diag}\left(\mathbf{D}_{nk}^{\mathrm{T}} \mathbf{D}_{nk}\right) \end{aligned}$$

   where $\text{diag}\mathbf{C}_k^{-1}$ is a diagonal matrix equal to the main diagonal of $\mathbf{C}_k^{-1}$:

$$\begin{aligned} \partial \text{AZ-}LL'/\partial \mathbf{C}_k^{-1} = \sum_n f(k|n) \left[\mathbf{C}_k - 0.5\text{diag}\mathbf{C}_k - \mathbf{D}_{nk}^{\mathrm{T}} \mathbf{D}_{nk}\right. \\ \left. + 0.5\text{diag}\left(\mathbf{D}_{nk}^{\mathrm{T}} \mathbf{D}_{nk}\right)\right] = 0 \end{aligned}$$

4. Prove that the solution obtained above (P1) solves this equation as well: take diag of (P1) and add to (P1). Thus, it turns out that ignoring the symmetry constraint for $\mathbf{C}$ did not change the result. But beware: in more complicated cases, this is not necessarily true. The identities for matrix derivatives used above can be found in Searle (1982).

Section 4.4

4.4–1 Derive Eqs. (4.4-13), (4.4-14), and (4.4-15) from MFT Eq. (4.4-7).

1. A_k estimation. Derive (4.4-7) for A_k,

$$dA_k/dt = \sum_\omega N_\omega f(k|\omega)/A_k + \lambda$$

   Consider this as an iterative discrete equation, so that $dA_k = A_k^{it} - A_k^{it-1}$, where A_k^{it} is a value of A_k at the iteration number it, and $(it-1)$ is the previous iteration. The right-hand side is evaluated at the $(it-1)$ iteration. Define $N_k = \Sigma_\omega N_\omega f(k|\omega)$. Choose $dt = A_k^{it-1}/N_k^{it-1}$ and $\lambda = -N$ and derive (4.4-13). Verify that the constraint on A_k is satisfied.

2. ω_k estimation. Derive (4.4-7) for ω_k,

$$d\omega_k/dt = \sum_\omega N_\omega f(k|\omega) \ (\omega - \omega_k)$$

   Show that this equation leads to (4.4-14), if you choose $dt = 1/N_k^{it-1}$. Write $d\omega_k = \omega_k^{it} - \omega_k^{it-1}$ and evaluate the right-hand side at $(it-1)$.

3. σ_k^2 estimation. Derive (4.4-7) for σ_k^2,

$$d\sigma_k^2/dt = \sum_\omega N_\omega f(k|\omega) \left[-0.5\sigma_k^{-2} + 0.5\sigma_k^{-4} (\omega - \omega_k)^2\right]$$

Show that this equation leads to (4.4-15), if you choose $dt = 2\left(\sigma_k^{it-1}\right)^4 / N_k^{it-1}$.

4.4–2 Demonstrate that at the maximum, $d\mathbf{S}_k/dt = 0$, (4.4-7) satisfies Eqs. (4.4-13), (4.4-14), and (4.4-15).

4.4–3 Derive Eqs. (4.4-13), (4.4-14), and (4.4-15) using a modification of the expectation-maximization (EM) algorithm. It is formulated as follows: an iterative maximization of AZ-similarity (4.4-6) can be achieved by maximizing at every iteration, it, the following expression:

$$\max_{\mathbf{S}_k^{it}} \left\{ \sum_{\omega} N_\omega f\left(k|\omega, \mathbf{S}^{it-1}\right) \ln\left[F\left(\omega|k, \mathbf{S}^{it}\right)/\hbar\omega\right] \right\}, \quad N_\omega = S(\omega)/\hbar\omega \tag{P2}$$

Here, fuzzy memberships $f(k, \mathbf{S}_k^{it-1}|\omega)$ are evaluated using the parameter values $\mathbf{S}^{it-1}$ estimated at the previous iteration, $(it-1)$. (A proof that this procedure always converges to max AZ-*LL* is considered in Section 4.6.)

1. Derive the A_k estimation equation. Account for the constraint (4.4-3) by using the method of Lagrange multipliers as in Problem 4.3-3.

   1.1. Maximization in (P2) leads to

   $$\sum_{\omega} N_\omega f\left(k|\omega, \mathbf{S}^{it-1}\right) \partial/\partial \mathbf{S}_k^{it} \ln\left[F\left(\omega|k, \mathbf{S}^{it}\right)/\hbar\omega\right] \Big\} = 0 \tag{P3}$$

   1.2. Evaluate the above for A_k (that is, substitute $\partial/\partial \mathbf{S}_k^{it} \rightarrow \partial/\partial A_k^{it}$). Obtain the following:

   $$\sum_{\omega} N_\omega f\left(k|\omega, \mathbf{S}^{it-1}\right)/A_k^{it} + \lambda = 0 \tag{P4}$$

   1.3. Multiply the above by A_k^{it} and sum over k. Obtain $\lambda = -\Sigma_\omega N_\omega = -N$ (compare to Problem 4.4-1).

   1.4. Combine this with (P4) and obtain

   $$A_k^{it} = \sum_{\omega} N_\omega f\left(k|\omega, \mathbf{S}^{it-1}\right)/N = N_k^{it-1}/N$$

   This leads to (4.4-13).

2. Derive the ω_k estimation equation using (P3). Evaluate (P3) for ω_k and obtain

   $$\sum_{\omega} N_\omega f(k|\omega, \mathbf{S}^{it-1})(\omega - \omega_k^{it}) = 0$$

   This leads to (4.4-14).

3. Derive the σ_k estimation equation using (P3). Evaluate (P3) for σ_k and obtain

   $$\sum_{\omega} N_\omega f\left(k|\omega, \mathbf{S}^{it-1}\right) \left[-\sigma_k^{it-1} + \sigma_k^{it-3}\left(\omega - \omega_k^{it}\right)^2\right] = 0$$

   This leads to (4.4-15).

4.4–4 Analyze conditions under which the ML equations maximizing the Einsteinian likelihood (4.4-7, 8) are equivalent to ME equations maximizing the entropy (4.4-29) subject to constraints (4.4-17).

1. Consider the Einsteinian Gaussian mixture models.
   1.1. Derive the ME equations for the model parameters maximizing entropy (4.4-19) subject to constraints (4.4-17). According to the method of Lagrange multipliers, maximization under the constraints is achieved by considering $\partial E'/\partial \mathbf{S}_k = 0$, instead of $\partial E/\partial \mathbf{S}_k = 0$, where $E' = \{E + \lambda[\varepsilon - N\Sigma_\omega\ F(\omega)] + \mu[(N - N\Sigma_\omega\ F(\omega)/\hbar\omega]\}$, and finding (λ, μ) that satisfy the constraints:

$$\partial E'/\partial \mathbf{S}_k = \sum_\omega\ [S(\omega)/\hbar\omega]\ \ f(k|\omega)\ \ [\partial \ln(F(\omega|k)/\partial \mathbf{S}_k] \\ - \lambda N \partial/\partial \mathbf{S}_k \sum_\omega\ F(\omega) - \mu N \partial/\partial \mathbf{S}_k \sum_\omega\ F(\omega)/\hbar\omega = 0$$

   1.2. Simplify constraints (4.4-17); substitute there the Einsteinian Gaussian mixture model, (4.4-9), and derive

$$N = N \sum_\omega\ F(\omega)/\hbar\omega = N \sum_\omega \sum_k\ A_k G(\omega|k) = N \sum_k\ A_k = N$$

   so this constraint is identically satisfied.

$$\varepsilon = N \sum_\omega\ F(\omega) = N \sum_\omega \sum_k\ \hbar\omega A_k G(\omega|k) = N \sum_k\ \hbar\omega_k A_k$$

   This constraint needs to be satisfied.

   1.3. Substitute 1.2 and the Einsteinian Gaussian mixture model, (4.4-9), into 1.1, and obtain the following equations for A_k and ω_k and σ_k^2:

$$\sum_\omega\ f(k|\omega) N_\omega / A_k - \lambda N \hbar\omega_k - \mu N = 0$$

$$\sum_\omega\ f(k|\omega) N_\omega\ (\omega - \omega_k)\ /\sigma_k^2 + \lambda N = 0$$

$$\sigma_k^2 = \sum_\omega\ f(k|\omega) N_\omega\ (\omega - \omega_k)^2\ / N_k$$

   The equation for σ_k^2 is exactly same as (4.4-15); the equations for A_k and ω_k matches (4.4-13), and (4.4-14), if we use $\lambda = 0$ and $\mu = 1$.

   1.4. It is only left to verify the constraint of energy conservation. Substitute the obtained expression for ω_k into the constraint and verify that it is satisfied identically.

2. Consider a general type model for $G(\omega|k)$ (non-Gaussian).
   2.1. Consider two model parameters A_k and ω_k (as a part of the set $\mathbf{S}_k$) defined as follows: $F(\omega|k) = \hbar\omega A_k G(\omega|k)$, where G is an arbitrary function satisfying $\Sigma_\omega\ G(\omega|k) = 1$; then, from $\Sigma_\omega\ F(\omega)/\hbar\omega = 1$, it follows that $\Sigma_\omega\ A_k = 1$. And define $\omega_k = \Sigma_\omega\ \omega G(\omega|k)$. This parameterization is consistent with the Einsteinian likelihood definition and does not impose any new restriction on $F(\omega|k)$ [which only have to satisfy the $F(\omega)$ normalization: $1 = \Sigma_\omega\ F(\omega) = \Sigma_\omega \Sigma_k\ F(\omega|k)$]. Therefore, the generality of our consideration is not limited by this parameterization. The purpose of these parameters is to capture the constraints (4.4-17) independently of any other parameters on which $F(\omega|k)$ might depend.

2.2. Rewrite the constraints (4.4-17) in terms of A_k and ω_k:

$$N = N \sum_{\omega} F(\omega)/\hbar\omega = N \sum_{\omega} \sum_{k} A_k G(\omega|k) = N \sum_{k} A_k = N$$

so this constraint is identically satisfied due to the normalization;

$$\varepsilon = N \sum_{\omega} F(\omega) = N \sum_{\omega} \sum_{k} \hbar\omega A_k G(\omega|k) = N \sum_{k} \hbar\omega_k A_k$$

We need to verify that this constraint is satisfied without changing the ML equations. The constraints (4.4-17) are entirely expressed in terms of parameters ω_k, A_k. Therefore, we need to verify the ML equations only for these two parameters. Equations for other parameters will match the maximum likelihood equations, similar to σ_k above.

2.3. Use the estimation equation derived above in 1.1 to estimate ω_k and A_k parameters:

$$\partial E'/\partial A_k = \sum_{\omega} N_\omega f(k|\omega)\ [1/A_k] - \lambda N \sum_{k} \hbar\omega_k - \mu N = 0$$

$$\partial E'/\partial \omega_k = \sum_{\omega} N_\omega f(k|\omega)\ (\partial/\partial\omega_k)\,[\ln G(\omega|k)] - \lambda N \hbar A_k = 0$$

2.4. If we set $\lambda = 0$, and $\mu = 1$, then parameters satisfy the ML equations. It only remains to verify the energy constraint. Let us write

$$(\partial/\partial\omega_k)\,[\ln G(\omega|k)] = (\omega - \omega_k)\,G'(\omega|k)$$

where G' is an arbitrary function; substitute this, along with $\lambda = 0$, into the equation for ω_k, and take a sum over k,

$$\sum_{k} \sum_{\omega} N_\omega f(k|\omega)\ (\omega - \omega_k)\,G'(\omega|k) = 0$$

2.5. If $G' = G'(k)$, it depends only on parameters $\mathbf{S}_k$ and does not depend on ω, the above equation is equivalent to

$$0 = \sum_{k} \sum_{\omega} N_\omega f(k|\omega)\ (\omega - \omega_k) = \sum_{\omega} N_\omega \omega - \sum_{k} N\ A_k \omega_k$$

And this is equivalent to the energy constraint.

3. Consider an example of non-Gaussian $G(\omega|k) = (1/\omega_k)\exp(-\omega/\omega_k)$, $\omega > 0$; $G(\omega|k) = 0$, $\omega < 0$. Verify that this model is of the type 2.5 above, therefore, for this type of model, the likelihood and entropy maximizations are equivalent.

4.4–5 Obtain Eq. (4.4-27). *Hints*: In the first line, inside { }: multiply and divide by pdf(X). In the second line, $I_{\mathrm{U},0|X,Y}$ by definition is the information about the universe given by (X, Y), when the IS and W are not interacting; in a noninteracting universe, pdf is a product of IS and W pdfs.

4.4–6 Compute $\Gamma_{\mathrm{W}|X,Y}$ (4.4-32). *Hints*: follow Section 4.4.2.

1. Compute the number of microstates in a macrostate $\mathbf{x}$, $\Gamma_{\mathbf{x}}$. Consider $N_{\mathbf{x}}$ photons in a macrostate $\mathbf{x}$; every one can be in $\Phi(\mathbf{S}, \mathbf{x})$ microstates; account for permutations of identical photons:

$$\Gamma_{\mathbf{x}} = \Phi(\mathbf{S}, \mathbf{x})^{N\mathbf{x}}/N_{\mathbf{x}}!$$

2. The total number of microstates compatible with the data and model, $\Gamma_{\mathrm{W}|X,Y} = \prod_{\mathbf{x}} \Gamma_{\mathbf{x}}$.

4.4–7 Prove that Eqs. (4.4-40) have the property that estimates are independent for statistically independent systems. *Hints*:

1. Consider two independent systems

$$\mathbf{x} = \mathbf{x}1, \quad N_{\mathbf{x}} = N_{\mathbf{x}1}, \quad F(\mathbf{S}, \mathbf{x}) = F(\mathbf{S}1, \mathbf{x}1), \quad \sum_{\mathbf{x}1} F(\mathbf{S}1, \mathbf{x}1) = 1$$

and

$$\mathbf{x} = \mathbf{x}2, \quad N_{\mathbf{x}} = N_{\mathbf{x}2}, \quad F(\mathbf{S}, \mathbf{x}) = F(\mathbf{S}2, \mathbf{x}2), \quad \sum_{\mathbf{x}2} F(\mathbf{S}2, \mathbf{x}2) = 1$$

and a joint system

$$\mathbf{x} = (\mathbf{x}1, \mathbf{x}2), \quad \mathbf{S} = (\mathbf{S}1, \mathbf{S}2), \quad N_{\mathbf{x}} = (N_{\mathbf{x}1}, N_{\mathbf{x}2}),$$
$$F(\mathbf{S}, \mathbf{x}) = F(\mathbf{S}1, \mathbf{x}1)F(\mathbf{S}2, \mathbf{x}2) \quad \sum_{\mathbf{x}1} F(\mathbf{S}1, \mathbf{x}1) = 1, \quad \sum_{\mathbf{x}2} F(\mathbf{S}2, \mathbf{x}2) = 1$$

2. Show that Eq. (4.4-40) leads to the same estimated values $\mathbf{S}1, \mathbf{S}2$ for the joint system as for each system considered separately.

2.1. Write the estimation equation for the $\mathbf{x}1$ system,

$$\max_{\mathbf{S}1}\left\{\sum_{\mathbf{x}1} N_{\mathbf{x}1} \ln F(\mathbf{S}1, \mathbf{x}1) + \lambda\left[\sum_{\mathbf{x}1} F(\mathbf{S}1, \mathbf{x}1) - 1\right]\right\}$$

Take $\partial/\partial\mathbf{S}1$ and obtain

$$\sum_{\mathbf{x}1} N_{\mathbf{x}1}(\partial/\partial\mathbf{S}1) \ln F(\mathbf{S}1, \mathbf{x}1) + \lambda\,(\partial/\partial\mathbf{S}1)F(\mathbf{S}1, \mathbf{x}1) = 0 \tag{P5}$$

2.2. Write the estimation equation for the joint system,

$$\max_{\mathbf{S}}\left\{\sum_{\mathbf{x}2} N_{\mathbf{x}2} \sum_{\mathbf{x}1} N_{\mathbf{x}1} \ln[F(\mathbf{S}1, \mathbf{x}1)F(\mathbf{S}2, \mathbf{x}2)]\right.$$
$$\left. +\lambda 1\left[\sum_{\mathbf{x}1} F(\mathbf{S}1, \mathbf{x}1) - 1\right] + \lambda 2\left[\sum_{\mathbf{x}2} F(\mathbf{S}2, \mathbf{x}2) - 1\right]\right\}$$

The gradient vector $\partial/\partial\mathbf{S}$ has two vector components, $\partial/\partial\mathbf{S} = (\partial/\partial\mathbf{S}1, \partial/\partial\mathbf{S}2)$; taking the first component, $\partial/\partial\mathbf{S}1$, leads to

$$\sum_{\mathbf{x}2} N_{\mathbf{x}2} \sum_{\mathbf{x}1} N_{\mathbf{x}1}(\partial/\partial\mathbf{S}1) \ln F(\mathbf{S}1, \mathbf{x}1) + \lambda 1\,(\partial/\partial\mathbf{S}1) \sum_{\mathbf{x}1} F(\mathbf{S}1, \mathbf{x}1) = 0$$

This equation for $\mathbf{S}1$ is equivalent to (P5) with $\lambda 1 = \lambda \Sigma_{\mathbf{x}2} N_{\mathbf{x}2}$.

2.3. Apply similar considerations to $\mathbf{S}2$. Show that using the similarity for equivalent photons leads to the same conclusion.

Section 4.6

4.6–1 Show that the last line in Eq. (4.6-7) is nonnegative. *Hint*: rewrite this line as

$$\sum_n \sum_k f(k|n,t)\,[\mathrm{ll}(k,n|t+dt) - \mathrm{ll}(k,n|t)] = \sum_k \sum_n f(k|n,t)$$
$$[d\,\mathrm{ll}(k,n|t)/d\mathbf{S}_k]\,d\mathbf{S}_k$$

Use Eq. (4.6-2) and obtain $d\mathbf{S}_k/dt = \Sigma_k\ \Sigma_n\ f(k|n,t)\ [d\,\mathrm{ll}(k,n|t)/d\mathbf{S}_k]$. Substitute this into the above expression and rewrite it as a sum (over k) of squares.

Section 4.7

4.7–1 Combine SLT and MFT: derive an SLT-like accurate "penalty" procedure for learning from small samples, which is applicable to compicated estimation problem characteristic of MFT. This is a complicated problem, appropriate for a doctoral thesis (or several).

MLANS: MAXIMUM LIKELIHOOD ADAPTIVE NEURAL SYSTEM FOR GROUPING AND RECOGNITION

MLANS is a neural network implementing MFT based on Bayesian similarity or likelihood, which was developed in the previous chapter. This chapter describes the development of statistical models for MLANS with relatively simple deterministic properties. These models characterize deterministic properties of every type of object or signal source by its mean value. Statistical properties are characterized by a probability distribution function (pdf). A class of objects may contain several types of objects, so the class pdf is a weighted sum of object-type pdfs. In statistics, such models are called mixtures. For example, when each object type is characterized by Gaussian pdf, the model is called a Gaussian mixture. Gaussian mixtures can model pdf of any shape, thus, statistical properties of mixture models are quite sophisticated. In addition to Gaussian mixtures, this chapter considers Wishart and Rician mixtures, which are appropriate for radar images. Architectures, neuronal equations, and learning procedures are described.

A number of issues are considered: grouping, clustering, and classification; supervised, unsupervised, and partially supervised learning; automatic learning of structure and complexity of MLANS, convergence properties, and various types of performance sensitivities. A large number of examples is considered, using real and synthetic data. This chapter discusses fundamental issues as well as details of implementation and application development.

5.1 Grouping, Classification, and Models

Classification and recognition refer to finding classes of data with a priori specified properties; grouping or clustering refers to finding natural regularities in data. Even so called natural regularities depend on measures of similarity, and every procedure for grouping or clustering specifies such measures explicitly or implicitly. In MLANS, measures of similarity are defined as likelihood, which is mathematically equivalent to distance measures in metric space that are adaptive and constrained by the structure of likelihood models. Similarity between MLANS models and data is adaptive and flexible; every MLANS agent-model

and the corresponding type of objects has its own adaptive metric, so that the overall metric is nonlinear and object-type dependent.

When detailed knowledge of the structure or dynamics of the observed objects is not available, recognition, classification, or grouping has to rely entirely on statistical properties of the data. MLANS original development was based on statistical models. Statistical models are designed to be flexible, adaptive, robust, and capable of accounting for various types of the available information. *Flexibility* means that any shape of the data probability distribution function (pdf) can be modeled resulting in a classifier of arbitrary complicated shape, as required by data. *Efficient adaptivity* means learning from a small amount of data. *Robustness* means that the performance is not too sensitive to the model assumptions. Available *information* may include labeled training data, unlabeled real-time data, real-time planned interactive sensing of the environment that might provide additional labeled or unlabeled data, knowledge of the pdf functional parametric form, some of the parameter values, or hints about object and data properties.

MLANS statistical models described in this chapter achieve these properties in the following way. *Flexibility* is achieved by modeling the data pdf as a superposition of basis functions forming a complete set in a functional pdf space. This ensures that any shape of pdf can be modeled. *Adaptivity* is achieved by modifying adaptive model parameters, and its efficiency is achieved by proper selection of the set of the basis pdfs, resulting in parsimonious utilization of model parameters and reduced neural network complexity. The maximum likelihood (ML) dynamics leads to fast learning from relatively few samples, reaching the fundamental mathematical performance bounds on speed of learning. *Robustness* is achieved by fusing all available sources of information within a hierarchical structure that combines adaptivity with real-time interactive environment sensing. These concepts are considered in this chapter in detail. We discuss theoretical and empirical issues of MLANS performance as well as practical issues of achieving adaptivity and robustness in applications.

Pdf models that use superpositions of several basis functions are called in statistical literature mixture models. These basis functions are also called components of a mixture; they may correspond to modes (local maxima) of the distribution functions, or to different types of objects within a class. If mixture components use a complete set of basis functions, any pdf can be expanded as a sum of these components. We refer to mixture components as types of objects or modes, or, in order to emphasize their adaptive dynamic nature, we call them agents.

Selection of appropriate functional forms for the individual components of a mixture, such as Gaussian, exponential, or uniform, is one way to account for a priori information. In addition to these general functional forms, a priori information can be used to increase efficiency of parameterization by imposing restrictions on model parameters. Some of the model parameters might be learned adaptively, while other parameters are fixed to a priori known values. In particular, specific structures can be imposed on covariance matrices. It is also important to utilize an appropriate number of mixture components and we will describe several approaches to accomplishing this. The MLANS architecture and learning rules for several types of mixture models are described. We consider Gaussian mixtures that are appropriate for most cases, when data contain random variabilities and when specific information on data properties is unavailable. We also describe MLANS utilizing mixtures of Wishart and Rician components that are appropriate for radar images due to specific physical mechanisms of scattering of electromagnetic waves at radar frequencies.

We first describe the MLANS statistical models for grouping or clustering. This is also called unsupervised classification, because no class labels are provided to the neural network. It is followed by a description of the traditional supervised classification, which assumes that a neural network first undergoes a training learning process, during which class labels are perfectly known. Then we consider more complex models for real-world applications, which often include imperfectly known class labels and a need to combine information from many observations without class labels with a few labeled observations. This comprises Sections 5.2 and 5.3.

Estimation of the optimal number of mixture components (or types of objects) within each class is but one aspect of finding the optimal complexity of the neural network. Section 5.4 discusses two approaches that MLANS utilizes for estimation of its optimal complexity: the maximum likelihood (ML) approach that optimizes the overall data characterization and the minimum classification entropy (MCE) approach that minimizes classification errors.

Several examples of MLANS classification are presented utilizing real sensory data as well as simulated data. While describing examples, we emphasize the fact that for many applications where systematic a priori information is not available, there are hints containing little bits and pieces of a priori information, which can be beneficial; in other cases, there are operational opportunities to acquire additional information in real time. We discuss such opportunities and the ways to utilize such additional information within MLANS.

5.2 Gaussian Mixture Model: Unsupervised Learning or Grouping

5.2.1 Architecture and Parameters

The architecture of MLANS based on a Gaussian mixture model is shown in Figs. 5.2-1 and 5.2-2. Figure 5.2-1 shows details of the parameter estimation subsystem, and Fig. 5.2-2 shows details of the data association subsystem. MLANS has as its input all the available data. These may include the data vectors of observations or features, $\{\mathbf{x}_n, n = 1, \ldots, N\}$, training information (labeled data set), and the results of interactive environment inspection. In case of unsupervised learning considered in this section, the available data are the observation vectors. Unsupervised learning is also called grouping or clustering, because we are looking for natural grouping of objects in the classification space. The components of the observation vectors $\mathbf{x}_n = \left\{x_n^i, i = 1, \ldots, D\right\}$ depend on the application: x_n^i could be a set of sensor measurements on an individual object such as intensities in various spectral bands, a set of pixel values within a certain region in an image, or a set of features extracted from measured data. The number of components, D, is called the dimensionality of the classification space. The output nodes of the parameter estimation subsystem contain the parameters of all classes and types of objects; we use indexes k and m to refer to classes and types. The output nodes of the association subsystem contain the weights associating each observation or input data sample with all classes and types of objects (agents).

We begin with a Gaussian mixture model, describing the probability distribution of each object-type by a Gaussian distribution. The architecture shown in Fig. 5.2-1 is suitable for such model. Parameters of Gaussian distributions are the mean vectors and covariance matrixes. The deterministic aspect of a Gaussian agent models the mean vector of the

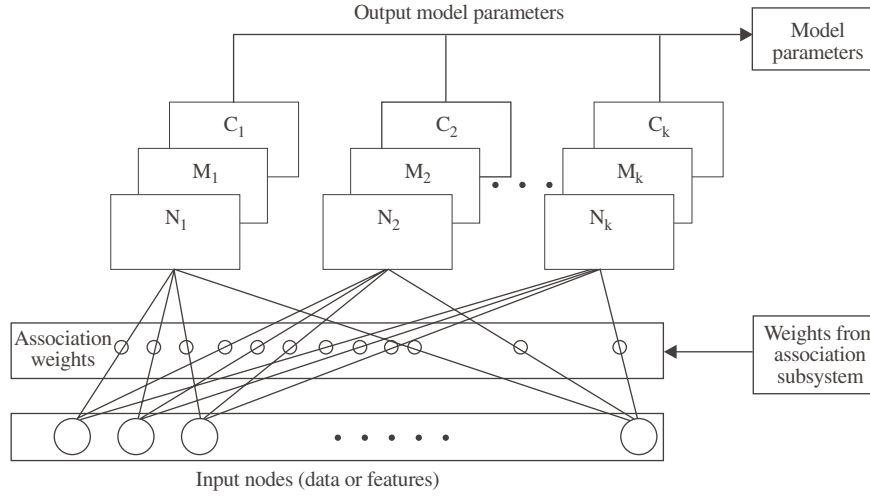


Figure 5.2-1 The MLANS modeling subsystem architecture.

object-type distribution, $\mathbf{M}_{km}$. The models are defined so that for some values of their parameters they match the expected values of the observations of this class and type objects,

$$\mathbf{M}_{km} = E\{\mathbf{x}_n|k,m\} \tag{5.2-1}$$

A statistical aspect of the agent is represented by the deviations of each observation $\mathbf{X}_n$ from the mean,

$$\mathbf{D}_{nkm} = \mathbf{x}_n - \mathbf{M}_{km} \tag{5.2-2}$$

The means describe predictable variability in the data, while the deviations are due to random causes, independent from one observation to another, such as most causes of sensor errors. The deviations are characterized by the covariance matrix:

$$\mathbf{C}_{km} = E\left\{\mathbf{D}_{nkm}\mathbf{D}_{nkm}^{\mathrm{T}}|k,m\right\} \tag{5.2-3}$$

here $\mathbf{D}_{nkm}^{\mathrm{T}}$ is a transposed vector-row, so that $\mathbf{D}_{nkm}\mathbf{D}_{nkm}^{\mathrm{T}}$ is an outer vector product, a matrix. Using vector component indexes explicitly, Eq. (5.2-3) is written as

$$C_{km}^{ij} = E\left\{D_{nkm}^{i}D_{nkm}^{j}|k,m\right\} \tag{5.2-4}$$

In addition to the mean and covariance parameters, a Gaussian mixture model is characterized by relative frequencies or rates of each object type:

$$r_{km} = N_{km}/N \tag{5.2-5}$$

where N_{km} is an expected number of (k,m)-type observations among the total number of N observations. In the statistical literature r_{km} are often called priors; to avoid confusion with prior or a priori information, we usually call r_{km} rates.

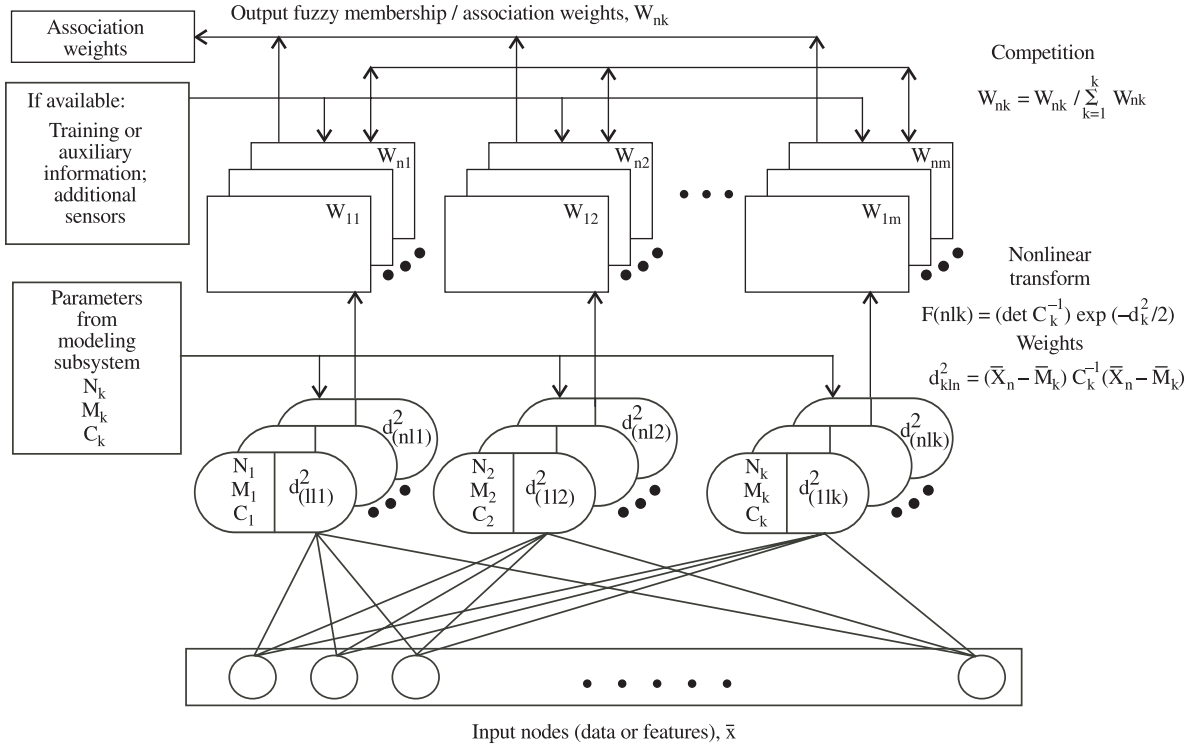


Figure 5.2-2 The MLANS association subsystem architecture.

5.2.2 Likelihood Structure and Learning Algorithm

5.2.2.1 Likelihood Structure

According to the above discussion of a Gaussian mixture model, the likelihood, L, or the total probability distribution function (pdf) for all observations $\{\mathbf{x}_n, n = 1, \ldots, N\}$ is a product of individual pdf $(\mathbf{X}_n)$, which are modeled as sums of Gaussian functions:

$$\text{AZ–}L = \text{pdf}\,\{\mathbf{x}_n, n = 1, \ldots, N\} = \prod_{n=1}^{N} \text{pdf}\,(\mathbf{x}_n)$$

$$\text{pdf}\,(\mathbf{X}_n) = \sum_{k=1}^{K} \sum_{m=1}^{M} r_{km}\ \text{pdf}\,(\mathbf{x}_n|k, m) \tag{5.2-6}$$

$$\text{pdf}\,(\mathbf{x}_n|k, m) = (2\pi)^{-D/2}\ (\det \mathbf{C}_{km})^{-1/2}\ \exp\left(0.5\, \mathbf{D}_{nkm}^{\mathrm{T}} \mathbf{C}_{km}^{-1}\, \mathbf{D}_{nkm}\right)$$

where $\det \mathbf{C}_{km}$ is the determinant of the matrix $\mathbf{C}_{km}$. The above model is a particular case of the general MLANS formulation given in Chapter 4 (Section 4.3). In the general case, the means, covariances, and rates are functions of model parameters; here, $\{\mathbf{M}_{km}, \mathbf{C}_{km}\}$ are the parameters of the model.

This model is appropriate for a wide variety of applications for the following reasons. Gaussian functions form a complete set of functions (this is easily proved by noticing that in the limit of $\mathbf{C} \to 0$, a Gaussian is a δ-function, see Problem 5.2-3). Therefore, a sum of Gaussians can model any pdf. This model is parsimonious, because the observations, $\mathbf{x}_n$, often include Gaussian noise, which tends to make each object-type nearly Gaussian. MLANS is robust with respect to small deviations of class distributions from Gaussian distributions, and large deviations from Gaussian shape are handled by additional object types. MLANS models deterministic variability by means of multiple object types, so that the remaining variability within each type is completely random and, thus, usually Gaussian.

Learning consists in estimating parameters of the model, the means, covariances, and rates of all modes, as well as computing association weights that associate data vectors with agent-modes, which accomplishes grouping or clustering of data. MLANS learning equations are designed to maximize the likelihood (5.2-6). The likelihood function (5.2-6) accounts for two sources of information: the structure of the probabilistic model and the data $\{\mathbf{x}_n\}$. In unsupervised learning, there is no distinction between classes and modes. We still will keep two indexes k and m, for later usage.

5.2.2.2 Modeling Subsystem

The output nodes of the modeling subsystem contain the estimated model parameters of all MLANS agents (classes and types of objects). Parameters are estimated by the corresponding ML neurons in Fig. 5.2-1, using the following neuronal equations (Problem 5.2-4 discusses the relationship of these equations to the general MLANS equations derived in Chapter 4). The estimated number of objects of each type

$$N_{km} = \sum_{n=1}^{N} f(k, m|n) \tag{5.2-7}$$

the estimated mean vector of each type

$$\mathbf{M}_{km} = \sum_{n=1}^{N} f(k, m|n)\ \mathbf{x}_n / N_{km} \tag{5.2-8}$$

and the covariance matrix of each type

$$\mathbf{C}_{km} = \sum_{n=1}^{N} f(k, m|n)\ \ (\mathbf{x}_n - \mathbf{M}_{km})^{\mathrm{T}}\ \ (\mathbf{x}_n - \mathbf{M}_{km})\ / N_{km} \tag{5.2-9}$$

Equations (5.2-7) through (5.2-9) implement the ML estimation of the parameters of the MLANS Gaussian mixture model, Eq. (5.2-6).

5.2.2.3 Association Subsystem

The MLANS association weights $f(k, m|n)$ were defined in Chapter 4 similar to the Bayes expression for a posteriori probabilities for an object n to belong to class k and type m:

$$f(k, m|n) = \text{pdf}\ (\mathbf{x}_n|k, m)\ / \sum_{k'm'} \text{pdf}\ \left(\mathbf{x}_n|k', m'\right) \tag{5.2-10}$$

This expression specifies the weight update at MLANS internal iterations, which is accomplished by an association-weight estimation subsystem shown in Fig. 5.2-2. Due to the normalizing denominator in (5.2-10), the weights update is a competitive learning: classes and modes "compete" for evidence to be associated with the object (or observation) n, while the total evidence for a given piece of data n to be in any class or mode is 1 (as given by the sum of all weights for a given datum).

Learning is accomplished by an iterative estimation of parameters and computation of weights. At the beginning of this iterative estimation, parameters are not equal to their true values and the weights $f(k, m|n)$ cannot be interpreted as a posteriori Bayes probabilities. The uncertainty of the object associations with classes and modes, which is represented in the weights, is of a more general nature than probability; weights also contain uncertainty due to unknown values of model parameters. The weights are fuzzy variables or membership functions.

In the process of learning, MLANS concurrently estimates parameters and weights, thus concurrently accomplishing association, while estimating model parameters. On convergence of this iterative learning procedure, the association weights become a posteriori Bayes probabilities and accomplish the optimal Bayes classification. Being probabilistic or fuzzy variables, weights accomplish fuzzy classification; the crisp, nonfuzzy classification can be obtained by classifying each object to the most probable class, or according to a suitable threshold determined by needs of specific applications.

Let us repeat that in the considered unsupervised classification, assignment of objects to the types is done on the basis of statistical properties of the data and statistical models. The classification of object types into classes requires additional information. Various examples of classification are discussed later.

5.2.2.4 Iterations and Convergence

During MLANS operations the weights are iteratively estimated according to (5.2-6) using the values of pdf model parameters estimated in the previous iteration, whereas the parameters are estimated according to (5.2-7) through (5.2-8) using the values of association weights estimated in the previous iteration. These internal iterations of MLANS continue until convergence.

The convergence is achieved when change between MLANS successive iterations becomes negligible. A natural definition of convergence for the ML estimation performed by MLANS could be in terms of the likelihood. However, testing likelihood changes for convergence is inconvenient, because the likelihood is a dimensioned value, depending on units used for feature measurements. Likelihood values change drastically between applications, and it is difficult to determine a priori the level of significant vs. insignificant changes. Alternatively, MLANS convergence can be defined in terms of probabilities or in terms of model parameters.

Our experience has shown that convergence in terms of MLANS model parameters leads to robust results: small variations in probabilities do not affect parameters significantly. (Small variations in parameters may affect probabilities for data close to a classifier boundary, therefore, convergence definition in terms of probabilities is less robust.) There are many possible ways to measure changes between two successive iterations of the MLANS models. We utilize the Bhattacharyya distance between pdfs at successive iterations for each mode, for the following reasons. The Bhattacharyya distance is a measure of dissimilarity

between two pdfs defined in terms of the overlap between the distributions and it is sensitive to differences in both the means (the centers of the distributions) and the covariances (the shapes of the distributions). The Bhattacharyya distance, B, between two Gaussian distributions can be expressed as a simple combination of their means and covariances:

$$B = \Delta\mathbf{M}\,\overline{\mathbf{C}}^{-1}\,\Delta\mathbf{M}/8 + 0.5\,\ln\left[\overline{\mathbf{C}}/\sqrt{\det(\mathbf{C}_1\mathbf{C}_2)}\right]$$
$$\Delta\mathbf{M} = (\mathbf{M}_2 - \mathbf{M}_1)\,; \quad \overline{\mathbf{C}} = (\mathbf{C}_1 + \mathbf{C}_2)\,/\,2 \tag{5.2-11}$$

and therefore it is easy to implement in MLANS. The Bhattacharyya distance B_{km} is computed between two successive iterations for each mode (k, m) and compared with a threshold. The convergence is defined as Bhattacharyya distances for all (k, m) being less than a threshold. Setting this threshold requires a little experience: too small a threshold will result in more iterations. To start, try threshold $= 0.001$.

The above equations completely specify the MLANS learning with two exceptions: first, MLANS initiation and second, determination of the number of MLANS agents or class types. MLANS initiation procedures will be discussed along with examples below. Determination of the number of MLANS modes or class types is related to the optimal complexity of the neural network and is discussed in Section 5.5.

5.2.3 Examples of MLANS Unsupervised Classification

5.2.3.1 Example 1: Real-World Application

Problem Description. In this example a quality control system has to identify a few defective parts among hundreds of perfect parts and other objects. It is expected that the approximate ratio of the numbers of (defective parts):(other objects):(perfect parts) would be on the order of 1:10:100 with significant variations. The specification was to reduce the necessary manual inspection rate to below 100 per 1000 objects. The classification has to be performed in a two-dimensional classification space of feature 1 and feature 2 extracted by a preprocessor from sensory data. These two features were the only sensory data available (sensor design and feature extraction, the two important aspects of the design of any recognition system, are not discussed here).

The actual data for a typical case are shown in Fig. 5.2-3 in the classification feature space. For the purpose of evaluation, the three classes of objects are shown here using three different symbols. The scatter in the distributions of each class in Fig. 5.2-3 is due to variations in aspect angles of the parts relative to the sensor and due to sensor errors.

Gaussian Data Characterization. In general, there is no reason to assume that the distributions for each class are Gaussian. It is still useful for an educational purpose, as a first step, to analyze class distributions under the Gaussian assumption. Gaussian distributions for each class can be estimated using standard estimation techniques, or equivalently, using MLANS with perfect supervision as described later in Section 5.3. We have estimated these Gaussian distributions using MLANS in a supervisory mode. For the increased statistical significance, additional measurements of class-1 and class-2 objects (the defective parts and the other objects) have been made, so that the Gaussian distributions were estimated using at least 100 objects from each class. We refer to these distributions as Gaussian-truth.

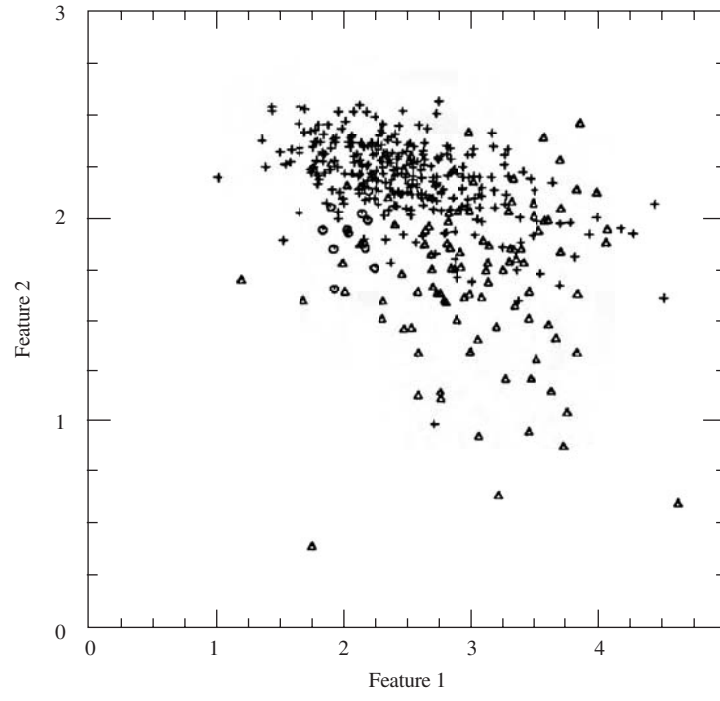


Figure 5.2-3 The distribution of measurements in the classification space, Example 1.

These distributions are shown in Fig. 5.2-4a. In this figure solid lines illustrate 2–σ concentration ellipses for each class; a 2–σ concentration ellipse is a standard convenient way of illustrating distributions; it is defined as a set of $\{\mathbf{x}_n\}$ at two standard deviations from the k, m-type mean according to

$$\mathbf{D}_{nkm}^{\mathrm{T}}\mathbf{C}_{km}^{-1}\,\mathbf{D}_{nkm} = 4 \tag{5.2-12}$$

The quantities in the left-hand side of this equation are deviations from class-type means and covariances as defined in Eqs. (5.2-6); considered geometrically, along any line crossing the center of the ellipse $\mathbf{M}_{km}$, the left-hand side quantity is a squared ratio of the distance between the center and the ellipse to the standard deviation along this line; correspondingly, the right-hand side is 2 squared.

Figure 5.2-4a also shows in a dotted line the Gaussian-truth-classifier boundary for the defected-parts-class 1 vs. the other two classes. This is calculated using the estimated Gaussian-truth distributions according to the standard definition of a likelihood ratio (*LR*) classifier given in Chapter 1:

$$\begin{aligned} LR &= \text{Likelihood(class1)/[Likelihood(class2)} + \text{Likelihood(class3)]} \\ &= \text{pdf}\,(\mathbf{x}_n|k=1, m=1) \,/\, \big[\text{pdf}\,(\mathbf{x}_n|k=2, m=1) \\ &+ \text{pdf}\,(\mathbf{x}_n|k=3, m=1)\big] = th \end{aligned} \tag{5.2-13}$$

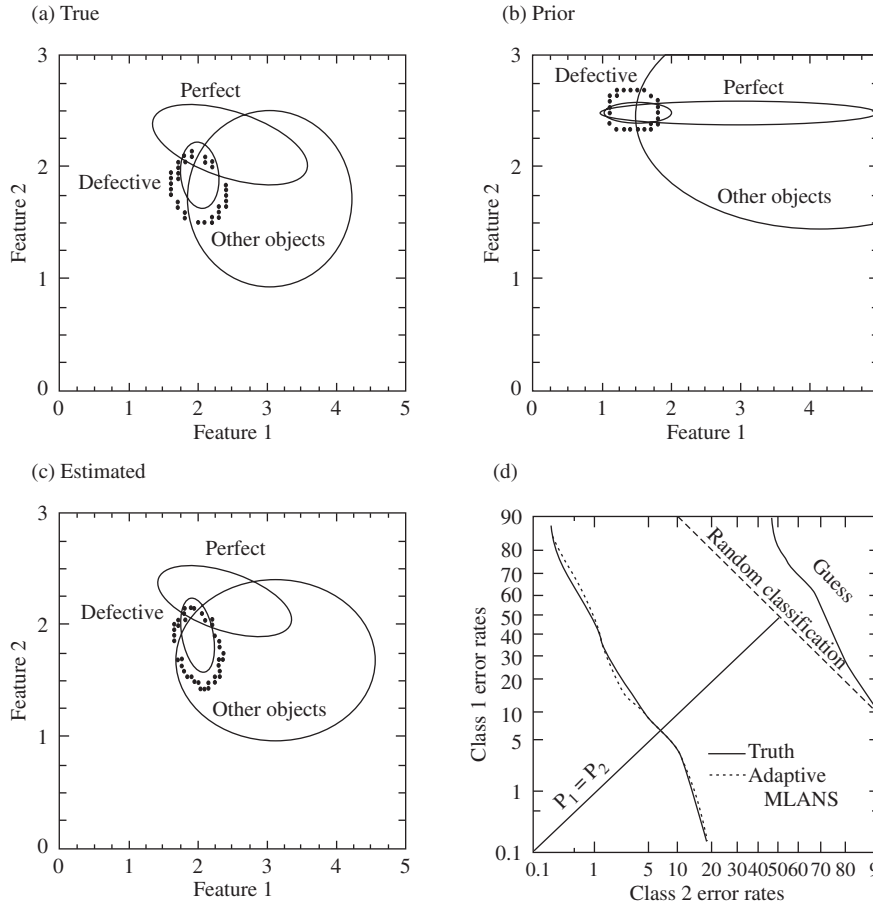


Figure 5.2-4 Example 1. (**a**) Gaussian-truth distributions corresponding to the data in Figure 5.2-3 and a classifier boundary; (**b**) initial guess distributions have no resemblance to the true ones; (**c**) MLANS estimated distrubtions are very similar to the true ones; (**d**) operating characteristic using adaptive classification is very similar to the one obtained with the Gaussian-truth distributions.

Here th is a threshold value; if LR $> th$, the nth object is classified to class 1. In Fig. 5.2-4a the classifier line was calculated for $th = 1$. An importance of the *LR* classifier is that when the distributions are accurately estimated, the *LR* classifier results in the minimal classification errors. The *LR* classifier can also be written in terms of a posteriori Bayes probabilities, Eq. (5.2-10); combining Eqs. (5.2-10) and (5.2-13):

$$P(k = 1, m = 1|n) = LR/(1 + LR) = th/(1 + th) \tag{5.2-14}$$

This equation is implemented in MLANS with ease, since the MLANS association weights, Eq. (5.2-10), are a posteriori Bayes probabilities, so that the *LR* classifier (5.2-13) with $th = 1$ is given by

$$W_{n11} = 0.5 \tag{5.2-15}$$

Unsupervised Learning. Let us remember that in this application unsupervised learning is required without the use of supervisory information on class assignments. Supervisory information is discussed above for evaluation and education purposes only and is not available for the actual MLANS learning. In our example, a typical case requires learning of class parameters and of recognition from a batch of approximately 1000 objects. A typical case data shown in Fig. 5.2-3 contain 816 objects. The 12 defective parts, if not marked, would be impossible to identify by eye as a separate cluster. Nevertheless, MLANS has been able to find the cluster of 12 defective parts; below we discuss the procedure, results, and an intuitive explanation for the internal working of MLANS.

Initiation of MLANS requires initial values of model parameters. If no information is available for an educated initial guess, it can be generated randomly. Robust initialization procedures are discussed later. Our experience has shown that usually MLANS is not sensitive to the initial guess, however, in some cases an educated guess results in faster convergence. We first discuss the MLANS learning using an educated initial guess. In this case, such a guess includes (1) using three MLANS modes, that is using one MLANS type for each class of objects, (2) using fixed (nonadaptive) class rates,

$$r_1 = 0.01, \quad r_2 = 0.1, \quad r_3 = 0.89 \tag{5.2-16}$$

and (3) using specific values of means and covariances for each class type as illustrated in Fig. 5.2-4b.

Comparing Fig. 5.2-4a and b, one can see that this educated guess is not a particularly good one: there are significant differences between the distributions in these figures. The a priori distributions for unsupervised learning are shown in Fig. 5.2-4b (the dotted line shows the corresponding classifier); they have been selected based on general considerations and they are far from the truth (3a). Nevertheless, the network has been able to find the cluster of 12 defective parts among 816 objects as well as the two other clusters. Classification of the three estimated clusters was based on the values of class rates fixed according to a priori information [Eq. (5.2-16)]. The results of unsupervised learning are shown in Fig. 5.2-4c. They are seen to be very close to the Gaussian-truth distributions in Fig. 5.2-4a.

The operating characteristics (OCs) are shown in Fig. 5.2-4d. An operating characteristic is a plot of the two types of errors as a function of the threshold in (5.2-13). The two types of errors in this case are probabilities of leakage and false alarm: P_{l} and P_{fa}. The probability of leakage is defined as the ratio of the missed defective parts to the total number of the defective parts, and it is related to another frequently used measure of performance, the probability of detection, P_{d}:

$$P_{\mathrm{l}} = 1 - P_{\mathrm{d}} \tag{5.2-17}$$

The probability of false alarm is defined as the ratio of the class-2 and class-3 objects misidentified as defective parts to the total number of the class-2 and class-3 objects. The OCs are obtained by counting leakages and false alarms for various values of the threshold th in Eq. (5.2-13). If only 12 defective parts were utilized for these computations, the resulting OCs would be very coarsely defined. Therefore, for the purpose of evaluation a larger number of objects was utilized. The axes in Fig. 5.2-4d are in the inverse-Gaussian scale.

Four OCs shown in Fig. 5.2-4d correspond to the following cases: random classification (coin toss), initial guess classifier (Fig. 5.2-4b), Gaussian-truth classifier (Fig. 5.2-4a), and adaptively learned classifier (Fig. 5.2-4c). If the initial guess classifier is used, the results are even worse than a random classification, indicating again that the guess was not a good one. Nevertheless, the adaptively estimated classification obtained with learned distributions is very similar to the Gaussian-truth classification obtained with perfect training. That these two curves are very similar indicates that MLANS in this case, without supervision and with only 12 samples from class 1, performs as well as if the perfectly labeled data in large numbers were available for training.

Because there are overlaps between the distributions, the defective parts cannot be identified without errors as indicated by OCs. The OCs are often characterized by the "diagonal" equal error point, at which leakage equals false alarm. In our case of the Bayesian *LR* classifier, the equal error is called the Bayes risk. It characterizes the inherent overlap of the distributions. As seen in Fig. 5.2-4d, the Bayes risk is about 5%. Based on the results of unsupervised learning, 80 objects were then selected for inspection and all 12 defective parts were correctly identified.

Discussion of Convergence Sensitivities. In this example, MLANS found a cluster of only 12 defective parts in the middle of 816 objects, while an eyeball examination of the object distribution in Fig. 5.2-3 does not indicate such clustering. Is it a lucky chance? How sensitive is this solution to various aspects of the initial guess and particular observation condition? To answer the first question, let us examine in detail the distribution in Fig. 5.2-3.

Fig. 5.2-5a shows a three-dimensional plot of the pdf estimated using the Parzen method. (The Parzen pdf estimation uses a sum of Gaussian components: one Gaussian component per observation, with the mean equals to the observation vector and with the same covariances selected to be somewhat smaller than the desired resolution, in our case it should be smaller than the covariance of class 1.) A little wiggle on the slope of the overall distribution in Fig. 5.2-5a corresponds to the class-1 cluster. This can be better seen in the bottom row, where 40 objects in class 1 are used for the illustration. Next, we subtract from this distribution the combined estimated distribution of classes 2 and 3. Figure 5.2-5b shows the remainder; now, the largest cluster corresponds to the class 1. Although this process is not a one-to-one description of the MLANS iterative estimation, it gives some intuitive idea of the information that is available for the estimation of the class-1 distribution. It can be seen that the successful result of the MLANS estimation is not a mere chance; on the other hand, the performance is close to a some sort of a fundamental limit: if there were fewer objects in class 1, the corresponding cluster would be indistinguishable from the ripples in Fig. 5.2-5b due to remaining random variabilities in class-2 and class-3 pdfs.

It would be useful to derive theoretically the probability that the ripples in Fig. 5.2-5b exceed the size of class-1 cluster (see Problem 5.2-12). Instead of an analytical solution, an alternative, numerical approach to this problem is described below.

Sensitivity to the Number of Class 1 Objects. We have varied the actual number of class-1 objects, without changes in the fixed rates [Eq. (5.2-16)]. The convergence was not significantly affected within the range between 8 and 20 objects in class 1. The effect of the increased number of objects in class 1 is illustrated in Fig. 5.2-5c and d. For 40 objects in class 1 and without changing numbers of objects in the other two classes, these figures

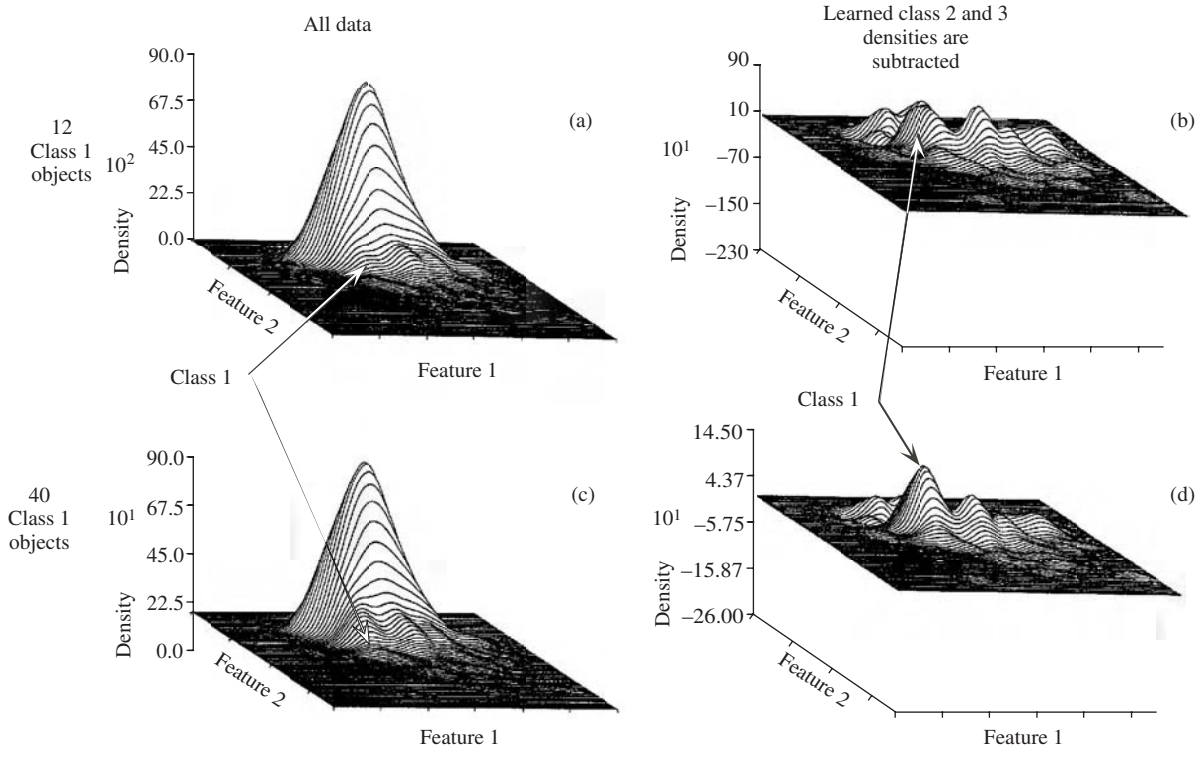


Figure 5.2-5 Robustness of learning in Example 1. The top row illustrates Example 1 (12 objects in class 1); the bottom row illustrates an effect of increased number of class-1 objects to 40; (**a**) and (**c**) Parzen estimation of the distribution; (**b**) and (**d**) the estimated class-2 and class-3 distrubtion is subtracted from (**a**) and (**b**), respectively.

show total pdf of all objects and the result of the subtraction of estimated class-2 and class-3 pdfs, similar to Fig. 5.2-5a and b: the corresponding class-1 cluster is more visible.

Sensitivity to the Initial Guess: Rates. One piece of prior information utilized in this example is an approximate ratio of class populations, or rates, fixed at 0.01, 0.1, 0.89 [Eq. (5.2-16)]. We found that the convergence is robust as long as the approximate ratio between classes is maintained. We varied class-1 rates between 0.005 and 0.04 and class-2 rates between 0.06 and 0.25 without much effect on the resulting classification. However, an attempt to estimate rates for all three classes led to an inaccurate estimation of the class-1 mean and covariance, and to an inaccurate classification. The bottom line here appears to be that in order to find a small class among classes with much higher populations, it is important to utilize the prior knowledge of the existence of a small class, even though the exact relative number of objects (the rate) in this small class is not known.

Sensitivity to the Initial Guess: Means and Covariances. Another piece of prior knowledge utilized was the initial guess for the means and covariances of the classes. As we have seen in the previous subsections, this initial guess was not a particularly good one. However, it did not prevent MLANS from converging to the correct distributions. The convergence

was not sensitive to the initial guess: we generated 10 random initial guesses and only once MLANS failed to converge to the same answer. In 9 of 10 cases, the number of internal iterations somewhat depended on the initial guess, varying between a few and a few tens; still the convergence results were insensitive to the initial guess. The case in which MLANS did not converge to the correct answer is studied further in Section 5.3. The approach in Section 5.3 is to combine unsupervised learning with real-time object inspection; this combined approach will be shown to result in a significant reduction of the number of required inspections: with a few additional inspections, correct convergence is obtained independent of the initial guess.

Conclusion and Further Exploration. The analysis of this example is still incomplete in that the estimation of the numbers of object-types or modes within each class was not addressed. In this example, one Gaussian mode has been used for each class of objects. We successfully identified the defective parts, and achieved an *adaptive* classifier performance similar to that of the classifier based on *supervised* training under Gaussian assumption. This is not a small achievement; however, the question remains if these results can be further improved by using more than three class types. How can the number of types to use be determined? From the point of view of the requirement to identify the defective parts, this problem could be considered as a two-class problem with unknown numbers of types within each class, which should be estimated from the data. This will be considered in Section 5.5.

5.2.3.2 Example 2: Parametric Characterization of MLANS Performance

Complicated classification problems are characterized by overlapping classes, large dimensionality, a large number of unknown parameters, and a small amount of data. In this example we evaluate the MLANS performance in 300 separate cases, characterized by different combinations of these complexity factors. We consider two unimodal classes with equal covariances in order to limit the number of different cases. MLANS, however, does not know that covariances are equal and estimates covariances for each class. When generalizing results of this example keep the following in mind. The performance in cases of multiple classes is usually limited by the two least separable ones. When separability is mostly due to the mean difference, the equal-covariance cases considered here give a good qualitative idea about achievable performance in case of unequal covariances.

The true overlap or separability between the classes we characterize by a k-factor:

$$k = \left[(\mathbf{M}_2 - \mathbf{M}_1)^{\mathrm{T}} \ \mathbf{C}^{-1} (\mathbf{M}_2 - \mathbf{M}_1) \right]^{-1/2} \tag{5.2-18}$$

where $\mathbf{M}_1$ and $\mathbf{M}_2$ are the means of two unimodal classes with equal covariances $\mathbf{C}$. In a case of classes with equal covariances a k-factor is an appropriate measure of separability; it is simply related to the Bhattacharyya distance (5.2-11),

$$k = [8B]^{-1/2} \tag{5.2-19}$$

and in a one-dimensional case, it measures the mean separability in units of the standard deviation, σ

$$k = (M_2 - M_1)/\sigma \tag{5.2-20}$$

The number of parameters per mode N_{par} is determined by the dimensionality D and by the parameters of the distribution that have to be learned. These parameters, defined in Eqs. (5.2-1) through (5.2-5), include rates r (one parameter per mode), means $\mathbf{M}$ (D parameters per mode), and covariances $\mathbf{C}$ [$D(D+1)/2$ parameters per mode]. Thus the total number of parameters per mode is

$$N_{\text{par}} = (D+1)(D+2)/2 \tag{5.2-21}$$

If rates are known a priori and means and covariances are to be estimated

$$N_{\text{par}} = D(D+3)/2 \tag{5.2-22}$$

If means are to be estimated alone

$$N_{\text{par}} = D \tag{5.2-23}$$

When the number of samples is too small, $N < N_{\text{par}}$, the problem is ill defined; on the other hand, when the number of samples is large, $N \gg N_{\text{par}}$, the problem is too easy. Also, low-dimensional highly separable problems are simple, but high-dimensional problems with low separability are complex. We evaluated the MLANS performance between these extreme conditions by selecting the combinations of parameters specified in Tables 5.2-1 and 5.2-2. Considered combinations of parameters result in $(5 \cdot 5 \cdot 3 \cdot 4) = 300$ cases. For each of the 300 cases, N data points per class have been generated using Gaussian distributions specified in Table 5.2-2.

The MLANS performance for these 300 cases is summarized in Fig. 5.2-6. In this figure MLANS performance is characterized by the Bhattacharyya distance, Eq. (5.2-11), between the estimated and the true distributions, averaged over the two classes in each case. This

TABLE 5.2-1
Values of Complexity Factors for Example 2

Complexity Factor	Values				
$k =$	0.5	1.0	1.5	2.0	3.0
$D = 2^x$, $x =$	0	1	2	3	4
$D =$	1	2	4	8	16
parameters	M		M, C		M, C, r
$N_{\text{par}} =$	D		$D(D+3)/2$		$(D+1)(D+2)/2$
$N = 3^y \cdot N_{\text{par}}$, $y =$	0	1	2		3
$N =$	N_{par}	$3 \cdot N_{\text{par}}$	$9 \cdot N_{\text{par}}$		$27 \cdot N_{\text{par}}$

TABLE 5.2-2
True Values of Gaussian Distribution Parameters for the Two Classes of Example 2

Rates	$r_1 = r_2 = 0.5$
Means	$\mathbf{M}_1 = (10, \ldots, 10)$; $\mathbf{M}_2 = (10 + k/D^{1/2}, \ldots, 10 + k/D^{1/2})$
Covariances	$\mathbf{C}_1 = \mathbf{C}_2 = \mathbf{1}$

Bhattacharyya distance is shown using gray scale; $B < 0.1$ represents a very good performance and $B \geq 0.5$ is a poor performance. Results in Fig. 5.2-6 are arranged in 15 cells by k-factor and parameters learned; within each cell, 20 cases are arranged by the dimensionality and the number of data samples. Since the number of data samples N is defined proportionately to the number of adaptive parameters N_{par}, the number of samples per parameter N/N_{par} is constant along each row and also for different rows with the same value of y.

Within each cell from top to bottom, there is a systematic improvement in performance due to increase in N/N_{par}. Within cells, horizontally, there is a tendency toward improvement in performance from left to right: although N/N_{par} stays constant, the total amount of available data, $N \cdot D$, grows with the dimensionality, so that $N \cdot D/N_{\text{par}}$ increases.

From cell to cell in the horizontal direction, one can see an improvement of performance for larger k-factors (larger separabilities). In the vertical direction, from cell to cell, both N/N_{par} and $N \cdot D/N_{\text{par}}$ stay constant, still there is some perceptible degradation with an increase in the number of unknown parameters: from M, to M, C, to M, C, r. This indicates that the increase in the number of adaptive parameters N_{par} is not completely counterbalanced by the proportional increase in the number of data samples available for learning N (and $N \cdot D$).

This example quantifies effects of complexity factors listed in Table 5.2-1 on the classification or clustering performance in terms of the accuracy of estimated pdfs. The MLANS performance is evaluated within a broad range of the complexity factors: k-factor $= 0.5$ to 3, $D = 1$ to 16, $N_{\text{par}} = 1$ to 136 per mode, and $N = 1$ to 3672 per mode. In addition to class separability measured by k-factor, the amount of data per unknown

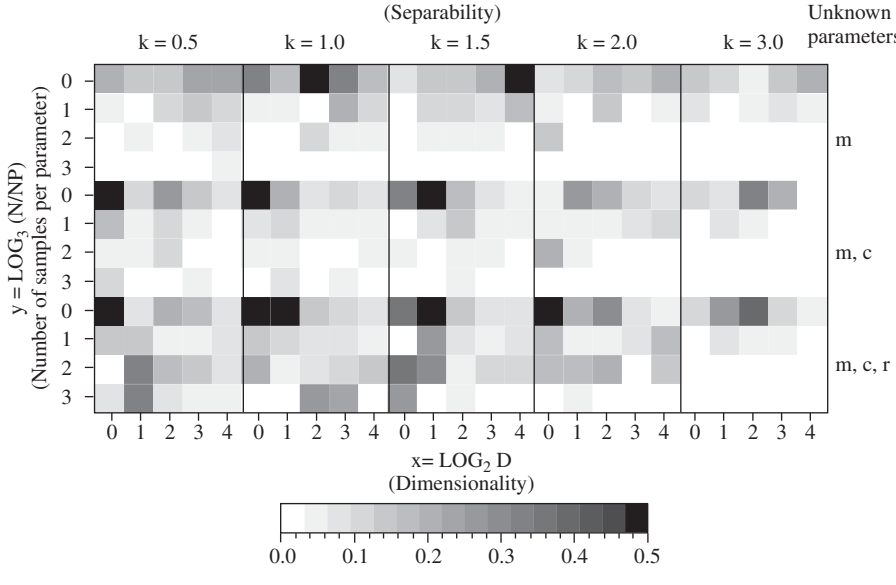


Figure 5.2-6 Example 2. Parametric characterization of the MLANS performance. Inherent separability is characterized by k-factor. MLANS performance is characterized by the Bhattacharyya distance between true and estimated distributions shown as gray scale.

parameter $N \cdot D/N_{\text{par}}$ is an important factor characterizing an overall complexity or difficulty of clustering. Still it does not capture all the aspects of complexity: the cases with the same values of $N \cdot D/N_{\text{par}}$ still show some degradation in the performance as the number of unknown parameters increases from M, to (M, C), to (M, C, r).

Characterization of all aspects of the classification complexity and of the MLANS performance does not seem feasible. The 300 cases considered in this example is a step toward characterizing some important aspects of this problem. Another approach to characterizing classification complexity is by studying fundamental bounds on performance, such as the Cramer–Rao bound. This is considered in Chapter 9.

5.2.3.3 Example 3: MLANS vs. Nearest Neighbor Comparison

This third classification example illustrates the MLANS performance using a standard data set studied in detail by Fukunaga (1972), and provides a comparison of the MLANS performance to ISODATA, a classical clustering algorithm based on the nearest neighbor concept (Fukunaga, 1972). As discussed in Chapter 2, most of classification algorithms and neural networks utilize the nearest neighbor concept, and their performances are expected to be similar to that of ISODATA.

In this example, unsupervised learning, or clustering is performed in eight-dimensional classification space. We consider cases with two and three classes. To facilitate evaluation, the data are simulated, so the data properties are exactly known. The true distribution of each class is Gaussian. These distributions are different in their means and covariances. Separability between classes is good but not perfect: the true Bayesian risk between each pair of the classes is about 2%. In Table 5.2-3 we summarize the results of clustering of two-class data for 100 objects in each class. MLANS significantly outperforms ISODATA: the misclassification errors obtained with MLANS are close to the Bayes risk, while the results obtained by Fukunaga with the ISODATA algorithm are significantly worse.

The fact that the classification errors are close to the Bayes risk, which is the minimal possible error if all parameters of class distributions are exactly known, suggests that MLANS yields an accurate estimation of all parameters of the distributions. This is further confirmed in Fig. 5.2-7, where some two-dimensional projections of the eight-dimensional distributions are shown: the estimated concentration ellipses are very close to the true ones, illustrating that the means and covariances are estimated very close to the true values.

Another way to quantify these results is by using the Bhattacharyya distances between the estimated and the true distributions. These distances are shown in Fig. 5.2-8 for each

TABLE 5.2-3
Comparison of the MLANS Neural Network and ISODATA Algorithm Using Standard Data Set; Two Classes, Eight-Dimensional Data

	MLANS		ISODATA Algorithm	
Actual	Assigned Class		Assigned Class	
Class	1	2	1	2
1	98	2	100	0
2	3	97	19	81

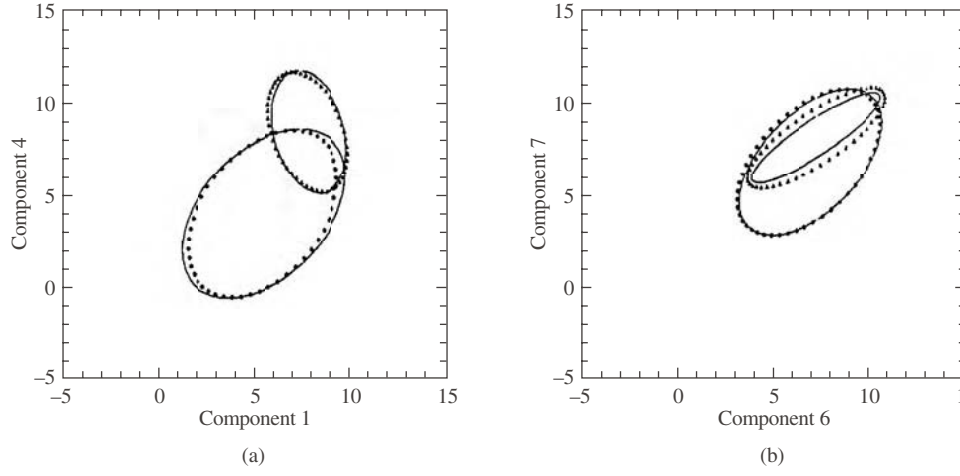


Figure 5.2-7 Examples of two-dimensional projections of eight-dimensional, two-class data, Example 3. Distributions are shown by illustrating $2-\sigma$ boundaries; solid lines are used for true distributions and symbols are used for the distributions estimated using the MLANS neural network.

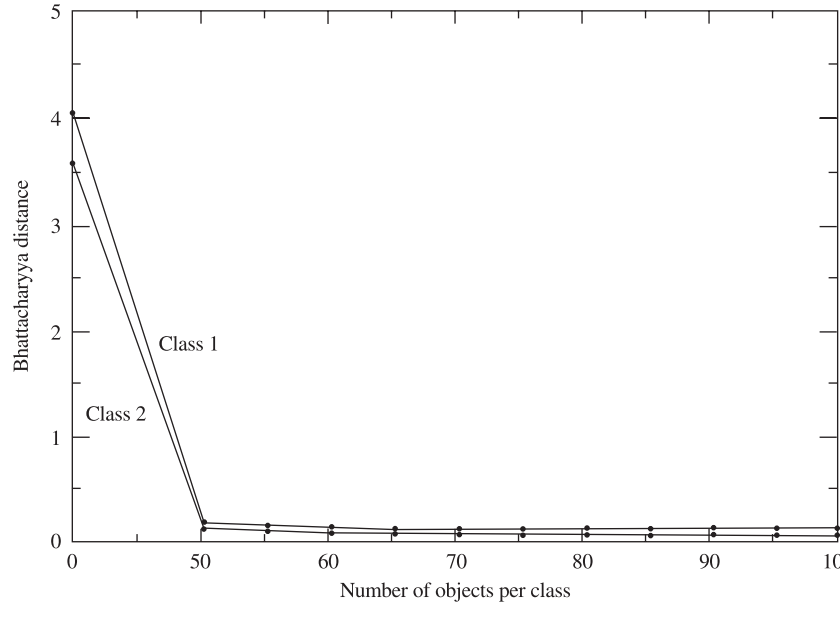


Figure 5.2-8 Bhattacharyya distances between the true and the estimated distributions for each class; unsupervised learning of two classes, Example 3. The initial guesses are very far from the true distributions; after adaptation a very close estimate is achieved with as little as 50 objects from each class.

class as a function of the number of objects in each class available for the MLANS learning process. In this figure, the initial Bhattacharyya distances between the true distribution and the initial guess for each class are quite large (about 4) due to the absence of prior knowledge of distributions. (In fact the "worst" initial guess was obtained by dividing the

entire region of the classification space occupied by all the data in halves along each axis, and by alternating the class assignments of these halves in such a way that the initial guesses are in the most distant eight-dimensional quadrants from the true distributions.) As MLANS starts learning, the Bhattacharyya distance is reduced to a very small number $B \sim 0.1$ to 0.2, with only 50 observations per class. In this eight-dimensional case, there are 45 independent parameters per class; so there are about nine scalar measurements per parameter, which is sufficient for an accurate estimation.

It is interesting to compare Example 3 with the previous Example 2 (Fig. 5.2-6). Example 3 for $N = 50$ objects per class is approximately comparable in difficulty of clustering to the rightmost bottom cell in Fig. 5.2-6, the case of $[k = 3;\ (M, C, r);\ x = 3; y = 0]$; see Problem 4.2–13 for details. Using the gray scale corresponding to this case, one can see in Fig. 5.2-6 that the Bhattacharyya is $B \sim 0.15$. Approximately the same Bhattacharyya is attained in Fig. 5.2-8 for $N = 50$. This illustrates that Example 2 results can be used for a general characterization of performance of the ML clustering.

There is always some difference between the estimated distribution and the true one, therefore the Bhattacharyya distance never equals zero, which would correspond to a perfect estimation. In other words, the Bhattacharyya distance is positively biased (in this specific sense), and this bias is a function of the number of observations available for learning. It follows from this discussion that there ought to be a fundamental limit to the minimal bias achievable with any algorithm or neural network, related to the limited amount of information contained in the finite number of observations. It turns out that such a limit indeed can be computed from information-theoretic considerations. This concept is further pursued in Chapter 9, where it is shown that the MLANS performance illustrated in Fig. 5.2-8 achieves this fundamental performance limit.

The internal convergence properties of MLANS learning process are illustrated in Fig. 5.2-9 by plotting the number of internal MLANS iteration cycles as a function of the number of objects. The initial number of iterations is small, ~ 10. Subsequent iterations are initialized when new objects become available to the network. These iterations provide only minor refinements to the previously obtained solutions so that one iteration is sufficient. The relatively large number of iterations in the beginning can be interpreted as a long relaxation time near the point of the phase transition (Kryukov, 1988); this interpretation is discussed later in Section 5.7.

Similarly good results are obtained with three classes. These results are summarized in Table 5.2-4. Again, classification errors obtained with MLANS are close to the Bayes risk, which is significantly better than the performance of the ISODATA algorithm. Some of the two-dimensional projections of these eight-dimensional results are shown in Fig. 5.2-10, and the Bhattacharyya distances between the estimated and the true distributions are shown in Fig. 5.2-11. Again, 50 observations per class are sufficient to obtain accurate estimates in this case.

As mentioned above, ISODATA is based on the nearest neighbor concept. This computational concept is utilized by most classification neural networks, therefore, as discussed in Chapter 2, their performances can be expected to be comparable to that of ISODATA. This was actually confirmed by testing other neural networks based on the nearest neighbor concept with this data set (Perlovsky, 1994a).

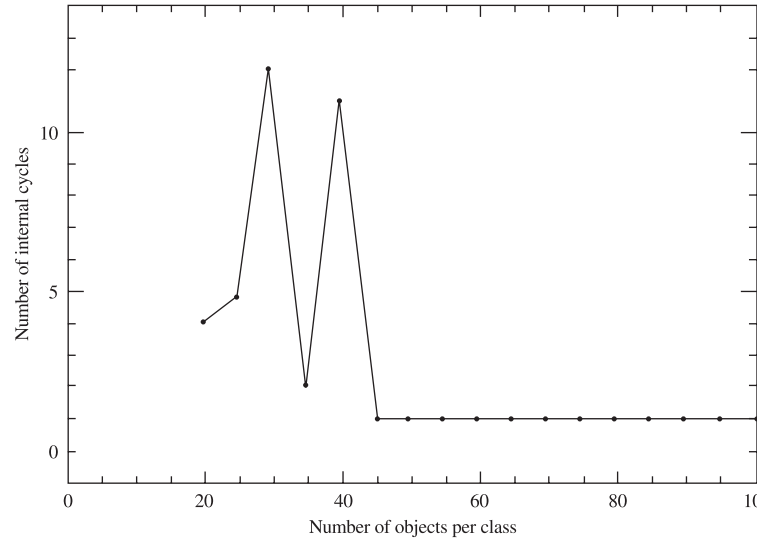


Figure 5.2-9 Number of internal iterations vs. number of objects per class (in Example 3). The network quickly converges with just one iteration for $N \geq 45$. For a small number of objects, which is insufficient for an accurate estimation of the eight-dimensional covariance matrices, transitions between maxima are evident, leading to metastable memory states, with the life-time an order of magnitude larger then the neuron-cylce time (∼one iteration).

TABLE 5.2-4
Comparison of the MLANS Neural Network and ISODATA Algorithm Using Standard Data Set; Three Classes, Eight-Dimensional Data, Example 3

	MLANS			ISODATA Algorithm		
Actual	Assigned Class			Assigned Class		
Class	1	2	3	1	2	3
1	98	2	0	98	2	0
2	2	97	1	27	73	0
3	1	1	98	18	0	82

5.3 Combined Supervised and Unsupervised Learning

5.3.1 Supervised and Unsupervised Learning

Learning is called supervised when first, a neural network or an algorithm is trained using data with perfect classification labels, which are said to be provided by a teacher. Subsequent applications or tests are performed using a different data set without further adaptation or learning. We will call supervision *perfect* if during the training, correct labels are provided

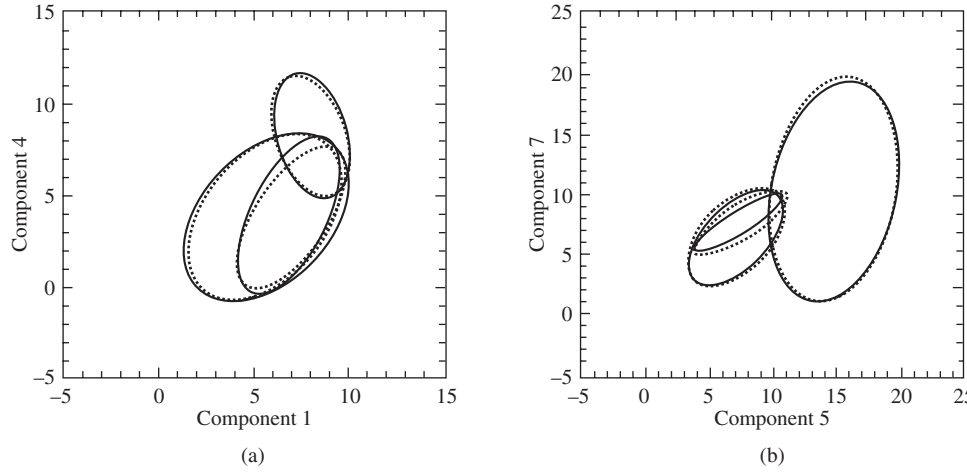


Figure 5.2-10 Examples of two-dimensional projections of eight-dimensional, three-class data, Example 3. Distributions are shown by illustrating $2-\sigma$ boundaries; solid lines are used for true distributions and symbols are used for the distributions estimated using the MLANS neural network.

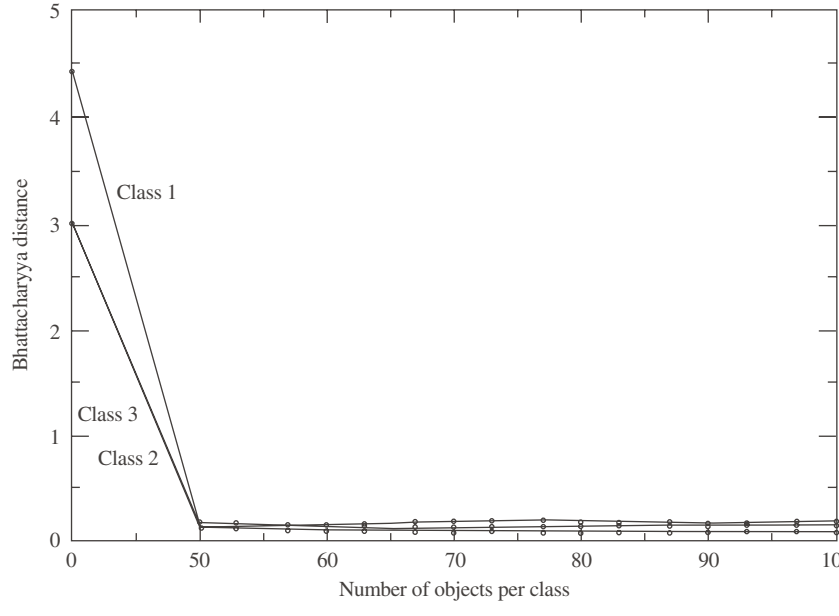


Figure 5.2-11 Bhattacharyya distances between the true and the estimated distributions for each class; unsupervised learning of three classes, Example 3. The initial guesses are very far from the true distributions; after adaptation very close estimates are achieved with as little as 50 objects from each class.

assigning *all* objects to *both classes and type-modes*. Other types of mixed unsupervised–supervised training include *class* supervision, when perfect *class* labels are provided for all objects without any *type* assignments; *partial* supervision, when only part of the objects are provided with class labels; and *imperfect* supervision when labels may be inaccurate. Partial or imperfect supervision often occurs in real-time applications. The term "real time" here refers not to any specific hardware implementation of MLANS, but is used in a sense similar to Carpenter (1989): the supervision in MLANS is accounted for via internal network dynamics.

5.3.2 Perfect Teacher

Perfect training is no different from the traditional Gaussian classifiers. In an important case of each type-mode distribution being truly Gaussian, such a process yields the optimal Bayesian classification, so it is desirable for a neural network to be able to reproduce such training. In MLANS this is achieved by supplying the teacher's labels directly to the output of the data association subsystem (Fig. 5.2-2) where the associating probabilistic weights are fixed to ones or zeroes according to the teacher's information. In this case, the entire data association subsystem is bypassed, and just single MLANS iteration is required for convergence. This type of MLANS operation was utilized for data characterization in Example 1 for computing the Gaussian-truth distributions.

5.3.3 Probabilistic or Fuzzy Teacher

The learning process of MLANS can be supervised, unsupervised, or a combination of unsupervised learning with partial or imperfect supervision. To achieve this flexibility the weights are modified to account for any available information concerning class or type assignments. If this information is probabilistic, association weights (on convergence) are a posteriori Bayesian probabilities, otherwise the weights are fuzzy variables approximating probabilities to some unknown degree. A teacher's assignment of a target n to a class k, type m will be denoted using a subscript T:

$$P_{\mathrm{T}}(k, m|n) \tag{5.3-1}$$

If a teacher provides perfect classification label, this value is 1 for a particular class and type, and zero for all the rest. In the case for which a teacher provides only probabilistic or fuzzy (or tentative) assignments, these values normally range between 0 and 1. This is not an essential requirement (though teacher's assignments have to be positive), because MLANS automatically normalizes teacher's assignments according to

$$P_{\mathrm{T}}(k, m|n) \rightarrow P_{\mathrm{T}}(k, m|n) / \sum_{k'm'} P_{\mathrm{T}}(k', m'|n) \tag{5.3-2}$$

so that they satisfy the constraint

$$\sum_{km} P_{\mathrm{T}}(k, m|n) = 1 \tag{5.3-3}$$

This is convenient in order to maintain a probabilistic interpretation of weights, which are modified according to

$$f(k, m|n) \rightarrow f(k, m|n) \cdot P_{\mathrm{T}}(k, m|n) \tag{5.3-4}$$

This modification results in the ML solution to the classification problem, which optimally fuses all the information from sensors and teachers (see Problem 5.3-1). The probabilistic interpretation of modified weights (5.3-4) also implies that teacher's information is statistically independent from the previously defined a posteriori probabilities $P(k, m|n)$, as is usually the case when a teacher "just knows" the class and type of an object, or derives its information from another sensor. In this case the modification (5.3-4) corresponds to the classical formula for combining independent probabilities.

5.3.4 Partial Supervision

A teacher's information may be incomplete, for example, when only a few objects are examined by a teacher. In this case of partial supervision, it is sufficient to modify weights $F(k, m|n)$ only for those objects n for which a teacher's information is available. This procedure results in an optimal (ML) fusion of all the available teacher's and sensory information; often a teacher's information on only a few objects leads to an improvement of classification of all objects resulting in a high learning efficiency, as illustrated in Section 5.3.5.1.

In an important case of partial supervision, a teacher provides information on class assignments only, $P_{\mathrm{T}}(k|n)$, and no information on the type. For example, when a neural network is trained to recognize several different objects viewed from different angles, the class of each object is known during training, however, it is desirable that the neural network determines on its own how many types it needs to represent each object adequately for robust and accurate classification, and also estimates parameters for each of these types. This type of training, with deterministic class assignments $P_{\mathrm{T}}(k|n) = \{1 \text{ or } 0\}$, supplied for all objects is the *most widely used class*-supervisory training. In a more general case, deterministic or probabilistic class assignments are available for only a few objects.

It turns out that for these cases, the weight modification formula (5.3-4) still provides an optimal solution that maximizes the likelihood of all the available information. This can be proven by appropriately modifying likelihood function and rederiving the ML estimation equations (see Problem 5.3-1). Below, we show this in a more illustrative way using probabilistic interpretation of weights. According to the rule of conditional probabilities, $P(k, m|n)$ can be represented as a product of two terms, the object's class probability $P(k|n)$ and the probability of an object's type, conditioned on an object's class:

$$P(k, m|n) = P(k|n) \cdot P(m|k, n) \tag{5.3-5}$$

A class probability, $P(k|n)$ is defined as

$$P(k|n) = \sum_{m=1}^{M} P(k, m|n) \tag{5.3-6}$$

It is modified using a probability supplied by a teacher, $P_{\mathrm{T}}(k|n)$, according to the rule of combining independent probabilities:

$$P(k|n) \rightarrow P(k|n) \cdot P_{\mathrm{T}}(k|n) \tag{5.3-7}$$

which yields the weight modification (5.3-4).

Weight modification [Eq. (5.3-4)] thus provides the ability to combine unsupervised and supervised learning. This is important in many practical cases, when large databases of unlabeled samples are available for training, while relatively few samples are labeled, because of time-consuming or expensive labeling procedures.

In many applications, environment can be interactively inspected during learning in real-time, by manual inspection or by an additional sensor. Often a few interactions during learning can significantly improve classification in complicated cases, when unsupervised classification is inadequate. Decisions as to which objects to inspect should be optimized in order to save the resources (e.g., time and sensor resources). Several ways to achieve this objective are discussed in Chapter 7, along with other issues concerning attention, multi-sensor fusion, and resource management. The partially supervised learning discussed in this section provides an efficient solution to one aspect of this problem, the evidence combining. As already mentioned, when teacher's assignments are deterministic or probabilistic, (5.3-4) leads to the optimal solution.

5.3.5 Examples

5.3.5.1 Example 1 Continuation: Combined Unsupervised and Interactive Learning

Here we continue consideration of Example 1 from Section 5.2.3, in particular, an unlucky initial guess, that converged to a wrong class-1 cluster (a local maximum of the likelihood). In such a case, when objects classified by MLANS as class 1 are inspected, the error becomes clear. At this point, the quality control system has an alternative: to proceed with the manual inspection of all the objects, or to reclassify all objects after the manual inspection of a few. This example demonstrates that there is much to be gained from the second approach of continuous adaptation. This can be done by using the partial supervision model described in the previous section.

In our experience, MLANS often found the cluster of defective parts right away; if it failed in the first attempt, it usually found the class-1 cluster after one or two inspections. Below, in order to construct a more complicated example requiring several inspection–reclassification cycles, we combine an unlucky initial guess with *not using* the a priori information about an approximate number of defective parts [Eq. (5.2-16)]. Still, the goal is to find all defective parts with fewer than 100 manual inspections, that is, by inspecting only about 10% of all the objects.

In this example, we inspect several most likely class-1 objects after every reclassification cycle. Two parameters should be defined to specify the procedure: how many objects are to be manually inspected in each inspection cycle, and when to stop. Both of these depend on operational aspects of specific applications: how expensive are the inspections, how long do they take, can they be performed in parallel, can errors be tolerated, etc. Therefore, an overall operation optimization is application specific. In addition, selecting the most likely class-1 objects is not necessarily optimal; this is discussed further in Chapter 7. Because of these difficulties, the optimal selection of objects for inspection (optimal attention mechanism) remains an unsolved problem. Nevertheless, efficient approaches can

be developed by combining ad-hoc practical considerations and simulations as discussed in this example.

Figure 5.3-1 illustrates the partially supervised procedure with 10 objects inspected at each cycle. We stop after 100 total inspections. It shows the numbers of defective parts (class-1 objects) identified in each cycle. There is significant improvement of performance due to continued learning with partial supervision vs. random search through the entire set of objects.

Example 4: Class Supervision

Here we illustrate a most widely used type of training: class-only supervision available for all objects. Two-dimensional feature distributions were simulated for two classes of objects, with three Gaussian types for each class. Parameters of these distributions are shown in Table 5.3-1. Figure 5.3-2a illustrates these distributions for all six types of objects, with vertical and horizontal axes corresponding to the two classification features; the true classifier boundary is shown by a dotted line. Figure 5.3-2b shows the results: distributions estimated from 50 observed objects.

Of these 50 observations there are only two objects of the third type of class 1; therefore, the covariance matrix for this object-type distribution, as estimated from the data, is numerically singular. To prevent such problems, MLANS tests determinants of covariances on each iteration. When a determinant falls below a predetermined threshold, several approaches can be utilized to remedy the situation. In this example, the sensor characteristics were known, and MLANS was instructed in such a case to set the covariance equal to the sensor noise covariance matrix. Despite this difficulty due to insufficient data,

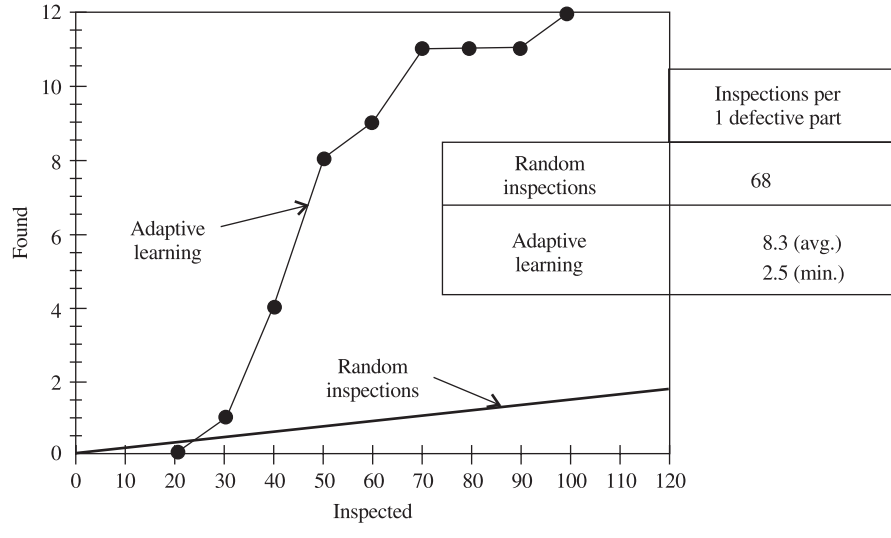


Figure 5.3-1 The number of defective parts found is shown as a function of the number of inspected parts for the modified Example 1. Adaptive learning with partial supervision results in significant improvement as compared with the random inspection.

TABLE 5.3-1
Parameters of Example 4 Distributions

Class	Type	Rates	Means		Covariances		
		r	M_1	M_2	σ_1	σ_2	ρ
1	1	0.25	3	9	0.5	1	0
1	2	0.20	10	6	0.5	1	0
1	3	0.05	17	16	0.5	1	0
2	1	0.20	3	12	1	0.5	0
2	2	0.20	12	6	1	0.5	0
2	3	0.10	17	13	1	0.5	0

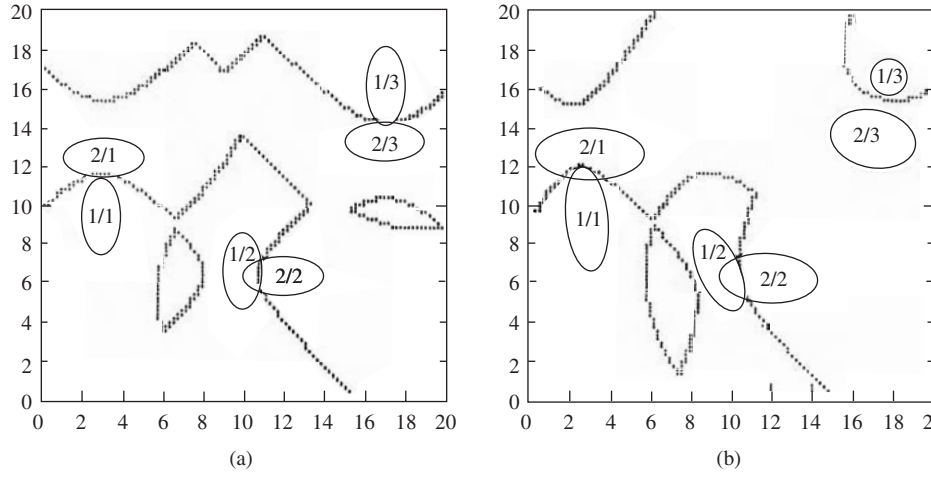


Figure 5.3-2 Learning a complicated shape of a classifier boundary (multiple modes) with class-only supervision, Example 4; (**a**) true distributions of the six types of objects of two classes; (**b**) MLANS estimated distrubtion after 50 observations.

the estimated classifier boundary is close to the true one and classification errors are close to the true minimum Bayesian errors.

5.4 Structure Estimation

5.4.1 Goals and Approaches of Structural Optimization: Models vs. Decisions

Examples considered above illustrate the neural network's ability to conceptualize, or to learn concepts: MLANS learns the concepts of object-types on its own. This learning is based on the a priori existence of these concepts as agents, which are built into MLANS'

statistical model. In previous examples, MLANS also knew the number of concepts to be learned. The conceptualizing ability is further enhanced in this section by learning the number of object-types or agent-concepts. The number of concepts is related to the neural network structure and complexity, because every agent-concept requires a dedicated set of neurons computing its model parameters.

An efficient (or optimal) structure of a neural network has to be learned from the available data. As more data become available, a more complicated architecture can be estimated. A most important parameter determining the structural complexity is the number of object types. Other complexity issues considered in this section include the structure of covariance matrixes for each type of objects, learning all parameters of the model vs. fixing some of them using a priori information, and restrictions imposed on adaptively learned parameters.

Estimation of the number of object-types for each class is similar to estimating the number of clusters in a clustering problem. This has been somewhat controversial in past clustering research. Many approaches have been suggested, which are useful for specific problems or rely on a priori experience or knowledge. For example, the vigilance parameter that controls formation of new clusters in the ART neural network is not determined by internal dynamics of the ART neural network and should be specified based on prior experience.

When determining the best number of clusters or object-types, it is important to keep in mind the goal, because the overall goal of the effort determines what is best. We have identified two general types of goals when solving classification and clustering problems. The first goal is to achieve the best characterization of the data for each class. The second goal is to make the best classification decision, or to achieve the smallest possible classification error. In a broader cognitive framework, the first goal is to improve the internal representation of the world, whereas the second one is to make the best decision concerning the real-world problem at hand. These two goals do not necessarily contradict each other: improving knowledge of the world in general leads to improved decisions and vice versa; solution of a particular recognition problem adds to the general knowledge of the world. However, these two goals do not coincide, first because of a competition for limited resources (such as sensors), and second, which is more subtle, a competition for interpreting (allocating, associating) evidence: the best characterization of the overall data may lead to different results than the best recognition of a single class. This is illustrated below.

5.4.1.1 Example 1: Continuation

In this example the number of object-types is not known a priori. In Section 5.2.3 we utilized three object-types: one type per class. However, since class distributions are not Gaussian, this choice is not necessarily optimal. Figure 5.4-1 illustrates results of using four and five modes for characterizing these data. We have kept the rate for one of the modes small, $r_1 = 0.01$, according to the discussion in Section 5.2.3 and Eq. (5.2-16), while other modes have adaptive rates estimated by N_{km} neurons according to Eqs. (5.2-7) and (5.2-5).

As could be expected, the additional modes in this figure concentrate around the center of the most populated class 2. Examination of MLANS weights [Eq. (5.2-10)] shows that these modes are mostly associated with class-2 objects. As discussed in the next section, these additional modes actually improved characterization of class 2 as compared to Fig. 5.2-2. Classes 1 and 3 are still characterized by a single Gaussian mode per class. But what is important for the classification decisions about class-1 objects, the characterization of class 1 was not improved: the estimated class-1 distributions in Fig. 5.4-1 are not as accurate as in

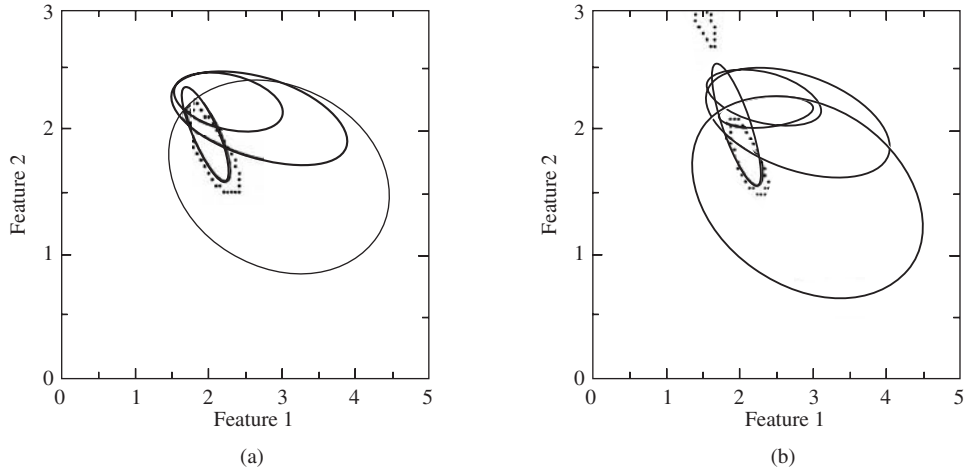


Figure 5.4-1 Example 1. Estimated distributions using four modes (**a**) and five modes (**b**). General learning can be in competition with solving a particular problem in hand.

Fig. 5.2-4c, when compared with the Gaussian truth of Fig. 5.2-4a. Similarly, classification performance of class-1 objects shown in Fig. 5.4-1 is not as good as in Fig. 5.2-4d. This example serves as a mathematical illustration of a well-known psychological phenomenon: learning in general and solving a particular problem could be in competition with each other. At the same time, some degree of general learning is necessary even to approach a solution of a particular problem at hand (in our example, estimation of class-2 and class-3 distributions is needed to find class 1).

The following sections discuss complexity estimation as reconciling the goals of general learning about the world and of decision making in a particular problem. We consider two fundamental quantities: likelihood and entropy. Maximization of likelihood is a data characterization method (the general learning about the world). Classification entropy is introduced to measure classification information, which is related to class separability. Minimization of classification entropy leads to improved classification decisions.

5.4.2 Maximum Likelihood Estimation of Structure

The ML approach is the most general and widely used approach to parameter estimation for the reasons discussed in Chapter 4, namely, the ML estimation is asymptotically unbiased and efficient. The ML approach, by definition, provides for a most likely characterization of all the available data. Utilization of the ML principle for the estimation of neural complexity including estimation of the number of object-types requires caution, because the likelihood itself is a biased estimator of the true likelihood value, when considered as a function of the variable number of parameters. Intuitively, this should be expected, because as more parameters are utilized in the model, a better fit might be expected. As discussed in Chapter 4, (4.7-1),

$$E\left\{LL\left(N, N_{\text{par}}\right)\right\} = LL_0(N) + N_{\text{par}}/2 \tag{5.4-1}$$

where LL is a log likelihood, N is a number of observations, N_{par} is a number of parameters, and $LL_0(N)$ is the expected value of the log likelihood for the true (not estimated) values of parameters, and $N_{\text{par}}/2$ is the bias. The log likelihood is proportional to the number of observations, so per observation

$$E\left\{LL\left(N, N_{\text{par}}\right)\right\}/N = LL_0(N)/N + N_{\text{par}}/(2N) \tag{5.4-2}$$

and the bias per observation, $N_{\text{par}}/(2N)$, asymptotically vanishes (at $N \to \infty$). However, for a finite number of observations, the bias in the likelihood often introduces significant bias into the estimated number of parameters. Therefore, we maximize the log likelihood corrected for the bias, which is called the Akaike Information Criterion, AIC, given by (4.7-2),

$$\text{AIC}/N = LL\left(N, N_{\text{par}}\right)/N - N_{\text{par}}/(2N) \tag{5.4-3}$$

For a clustering problem with M agents, the total number of parameters [Eqs. (5.2-1), (5.2-3), and (5.2-5)] is

$$N_{\text{par}} = M^*(D+1)(D+2)/2 \tag{5.4-4}$$

And an extension of the Akaike-type modification of the ML principle for clustering is

$$\text{AIC} = LL\left(N, N_{\text{par}}\right)/N - M^*(D+1)(D+2)/(4N) \tag{5.4-5}$$

Maximizing this expression has the attraction of providing an asymptotically unbiased and efficient estimate of a number of object-types, which entirely relies on the internal dynamics of the MLANS and does not need any prior knowledge or experience. We would refer to this modification of the ML as AIC, or simply the ML.

An application of AIC to the Example 1 problem is illustrated in Fig. 5.4-2. In this figure AIC is shown for $M = 3$, 4, and 5, and it peaks at $M = 4$, so that the ML chooses four object-types as the best characterization of the distribution of these data. This four-mode distribution characterization has been shown in Fig. 5.4-1. This example illustrates the point that learning the best overall model of the world does not necessarily lead to the best solution of a particular problem at hand: classification of a small class.

Mathematically, this result is explained as follows. AIC (as well as a likelihood function) is a homogeneous quantity that is maximized in a uniform way over all observations. Maximizing AIC is not satisfactory in cases in which there is particular interest in finding (with minimal error) a class containing a small number of objects among other classes containing large numbers of objects. The ML and AIC approaches could be, so to speak, "biased" toward a general learning of the world (an accurate description of large classes) *at the expense* of solving the problem in hand (finding the small class). A mitigation of this circumstance is discussed in the next section.

5.4.3 Minimum Classification Entropy

The ML approach to structural optimization considered above as well as other data characterization methods used to find the number of cluster do not address directly a goal of

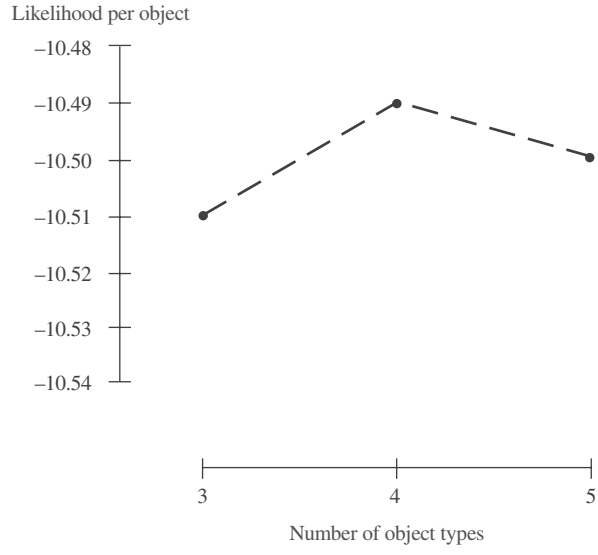


Figure 5.4-2 Learning the number of object types using the maximum likelihood (AIC). The best overall model of the world does not necessarily lead to the best solution of a particular problem at hand: classification of a small class.

minimizing classification errors, which is the specific problem to be solved. This section discusses a method that is more directly related to the minimization of classification errors. A straightforward minimization of a classification error for finding the best neural complexity, including the best number of object-types or clusters, is a possible approach for supervised learning, although a cumbersome one due to the complicated procedures required to estimate classification errors. In case of unsupervised distribution, the error is unknown, and the decision should be based on other principles. Such a principle is described in this section, based on minimization of classification entropy. The intuitive basis for this procedure is in maximizing the crispness of fuzzy classification, or equivalently, in maximizing class separability, and it closely approximates minimization of classification errors.

For the purpose of minimizing classification errors, a classification entropy (CE) of the data set $\{\mathbf{x}_n\}$ with regard to a set of a posteriori Bayes classification probabilities $\{P(k|n)\}$ (5.2-10) is defined as a mean value of the logarithm of the classification probability averaged over all classes, and it is estimated as follows:

$$\mathrm{CE} = -\sum_{n=1}^{N}\sum_{k=1}^{K} P(k|n)\ \ln\ P(k|n) \tag{5.4-6}$$

Classification entropy is a negative measure of classification information contained in a set of probabilities $\{P(k|n)\}$. CE is a nonnegative quantity that reaches its minimum $E = 0$ when classification is nonfuzzy, that is, when each object is assigned to a single class with a probability 1, so that all $P(k|n) = 0$ or 1. Because weights $f(k, m|n)$ estimate a posteriori Bayes probabilities $P(k, m|n)$, the Minimum Classification Entropy (MCE) neuron computes CE as follows:

$$\mathrm{CE} = -\sum_{n=1}^{N}\sum_{k=1}^{K}\left[\sum_{m=1}^{M} f(k, m|n)\right]\ \ln\left[\sum_{m=1}^{M} f(k, m|n)\right] \tag{5.4-7}$$

Beginning with preset values of number of classes K and number of types for each class M_k, the MCE neuron calculates CE (5.4-7), then it resets values of K and M_k and reinitiates MLANS. The resulting new value of CE is compared with the previous one and iterations continue until the minimum value of CE is found. In many applications, the number of classes K is predetermined by the problem formulation, and only the numbers of types for each class (M_k) should be learned from data. When the number of different M_k to be considered is small, an exhaustive search is appropriate, otherwise gradient descent can be used (that is, the search continues only in the direction of the gradient).

5.4.3.1 Example 1: Continuation

An application of the MCE principle in this example is illustrated in Fig. 5.4-3, were the CE values are shown for $M = 3, 4$, and 5 (since this is unsupervised learning, only the total number of object types, M, is defined). It reaches minimum at 3, indicating that the MCE results in choosing three object-types as the best for these data. As previously discussed, this choice results in the minimal classification error. Thus, MCE results in a better solution of the problem of finding a small class, as compared to the ML principle that resulted in a better overall internal "world" representation. The number of types in this problem should be selected using the MCE principle.

5.4.4 Other Structural Issues

This section considers several techniques for limiting the complexity of the adaptive structure of MLANS by restricting the number of independent adaptive parameters in the MLANS statistical model. This keeps up with the general model-based approach of utilizing known phenomenology to construct adequately flexible models without excessive numbers of adaptive parameters.

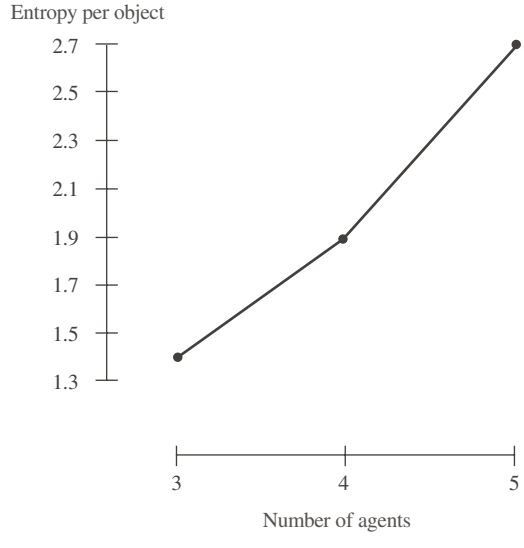


Figure 5.4-3 Learning the number of object types or MLANS agents using minimum classification entropy results in a better solution than AIC ML criterion for the particular problem of finding a small class.

5.4.4.1 Fixed Parameters

A straightforward and simple technique is to make an a priori decision as to which parameters should be adaptively learned and which should be fixed using a priori knowledge of the problem phenomenology. We used this technique already in Example 1, where object-type rates have been fixed using a priori knowledge. Similarly, any component of the mean vectors for some modes can be fixed. In Example 4, Section 5.3.5, the covariance matrix was fixed for a mode with fewer than a predetermined number of data points. Imposing more complicated constraints on covariance matrices is nontrivial and some of the techniques are considered below.

Constraining covariance matrices can be desirable, especially in high-dimensional cases, because the number of parameters in a covariance matrix grows as the second power of the dimensionality. This can be done either based on a priori knowledge of the phenomenology of the process, or in order to improve robustness if the number of observations is insufficient to estimate a covariance matrix of a particular object-type. Several simple alternatives are possible. The covariance matrix can be set to a predetermined (small) value, such as an observation or sensor noise covariance matrix, or to a large value, such as the overall covariance matrix of all the objects, depending on the objective and prior knowledge. Or a sensor noise covariance can be added to the estimated covariance.

5.4.4.2 Structured Covariances

Imposing a structure on a covariance matrix should not be done in an ad-hoc way: for example, restricting a covariance matrix to a desired structure by setting to zero nondesirable matrix components may lead to a nonpositive definite matrix, which is not suitable as a covariance estimate and may be noninvertible. Instead, the simplest approach is to use a diagonal covariance matrix. With a diagonal covariance, it is only necessary to make sure that the diagonal elements are not too small numerically. This can be achieved by setting a predetermined threshold, C_0:

$$C_{ii} \rightarrow \max\left[C_{ii}, C_0\right] \tag{5.4-8}$$

or by adding a small diagonal matrix:

$$C_{ii} \rightarrow C_{ii} + C_0 \tag{5.4-9}$$

Both these procedures are always numerically safe.

However, a simple diagonal covariance does not account for correlations between features and may be inadequate in that important classification information, if contained in the correlations, will be missed. A band-limited covariance structure,

$$C_{ij} = 0, \quad \text{for } |i - j| \leq b \tag{5.4-10}$$

can be used to account for some of the correlations. As mentioned, a band-limited covariance matrix should not be estimated by fixing out-of-band elements to zero; this may result in a nonpositive definite matrix. A simple approach, however, can be used to properly condition such an estimation. A matrix is called diagonally dominant, if for each row i

$$C_{ii}^2 \geq \sum_{j \neq 1}^{J} C_{ij}^2 \tag{5.4-11}$$

Further, a diagonally dominant matrix with positive diagonal elements is positive definite. Therefore, to ensure positive definiteness of an estimated structured covariance, estimated covariance in-band components can be modified in such a way that the condition (5.4-11) is satisfied. For example, off-diagonal elements can be modified as follows:

$$C_{ij} \rightarrow C_{ij} \cdot \exp\{-\alpha \cdot |i - j|\}, \quad \alpha > \ln 2 \tag{5.4-12}$$

Another appropriate modification is

$$C_{ij} \rightarrow C_{ij}/|i - j| \tag{5.4-13}$$

Similar types of modification can be used to condition not only band-limited matrices, but any type of imposed structure. A disadvantage of these ad-hoc techniques is that some information contained in the correlations is lost. A more advantageous approach to estimating structured covariances while preserving important information can be based on the estimation of Choleski factors of covariance matrices (Perlovsky and Marzetta, 1992).

5.5 Wishart and Rician Mixture Models for Radar Image Classification

In some applications non-Gaussian mixture models are more appropriate than Gaussian ones because of the nature or the physical sources of signals. This section considers radar measurements containing phase as well as amplitude. The characterization of radar signals is accomplished using mixture models of complex Gaussians, Wishart, and Rician density functions.

5.5.1 Synthetic Aperture Radar

A typical operation of a synthetic aperture radar (SAR) mounted on an aircraft can be described in a simplified way as follows. The radar transmits electromagnetic wave pulses that are relatively short and wide-angle ($\sim$ a few tens of degrees), perpendicular to the line of flight:

$$\mathrm{Re}\left[a(t) \exp\left(-2\pi i f_0 t\right)\right] \tag{5.5-1}$$

where $a(t)$ and f_0 are the known modulation amplitude and carrier frequency. A reflector, such as a feature on the ground, a tree, or a searched object, returns a reflected pulse

$$\mathrm{Re}\left\{k_n \cdot a\left(t - \tau_n\right) \exp\left[-2\pi i f_n \left(t - \tau_n\right)\right]\right\} \tag{5.5-2}$$

Here k_n is the reflection coefficient and f_n and τ_n are the frequency and time delay of the returned pulse. The frequency f_n is different from f_0 due to the Doppler shift determined by the relative reflector-to-radar velocity along the line of sight v,

$$f_n = f_0\,(1 + 2v/c) \tag{5.5-3}$$

where c is the speed of light. The time delay τ_n is determined by the range R_n to the reflector n:

$$\tau_n = 2R_n/c \tag{5.5-4}$$

Usually there are thousands of reflectors contributing to each return pulse, $S(t)$:

$$S(t) = \mathrm{Re}\sum_{n=1}^{N}\{k_n \cdot a\,(t - \tau_n)\exp\left[-2\pi i f_n\,(t - \tau_n)\right]\} \tag{5.5-5}$$

Reflectors at different ranges are resolved by removing the known shape of the modulation signal $a(t)$ from the returned pulse. The accuracy of this is determined by the radar bandwidth which is very high, permitting the resolution to a few feet or better. The resolution along the line of flight, called the cross-range resolution, is accomplished by utilizing the Doppler frequency shift in the returned pulse. To achieve high resolution and unambiguous resolution of range and cross-range reflectors, many hundreds to thousands of SAR pulses are coherently processed. Coherent processing implies that the relative phases of the pulses are accurately accounted for as if the pulses are transmitted and received by a very large aperture antenna, hence the name *synthetic* aperture.

During processing of the received pulse, the carrier frequency is removed by the following operations:

$$\begin{aligned} I(t) &= S(t)\cdot\cos(2\pi i f_0 t) \\ Q(t) &= S(t)\cdot\sin(2\pi i f_0 t) \end{aligned} \tag{5.5-6}$$

This defines an in-phase I and quadrature Q components of the signal; both are needed to preserve the amplitude and the phase of the return signal $S(t)$. These two components form a complex processed signal

$$I(t) + iQ(t) \tag{5.5-7}$$

After processing, the data obtained from an SAR may be displayed as a two-dimensional image of the area being illuminated by the radar beam. To preserve all the information contained in the radar return, every pixel must be characterized by a complex value.

5.5.2 Data Description

Most of examples illustrated in this section are taken from the analysis of data obtained from a polarimetric SAR, employed by NASA in its search and rescue mission, when looking for small downed aircraft. It transmits two different polarizations of electromagnetic waves and for each transmitted polarization the two polarizations are received. Altogether, four signals are received, so that four complex images are available for performing detection and identification, $s_{\mathrm{vv}}, s_{\mathrm{vh}}, s_{\mathrm{hv}}, s_{\mathrm{hh}}$. Here s_{vv} and s_{hh} are the two copolarized returns (vertical or horizontal transmitted and the same received) and s_{vh} and s_{hv} are the cross-polarized returns. It is usually assumed that $s_{\mathrm{vh}} = s_{\mathrm{hv}}$. In the examples considered below, the

cross-polarized return is actually computed using both the s_{vh} and the s_{hv} returns and performing phase equalization to account for the fact that the phase in the SAR images might not be properly calibrated. The radar return of the nth pixel, $\mathbf{s}_n$, may therefore be written as a three-dimensional complex vector

$$\mathbf{s}_n = [s_{vv}, s_{vh}, s_{hh}]^T \tag{5.5-8}$$

The data provided by the NASA Goddard Space Flight Center have been recorded as a pixel covariance $\mathbf{K}_n$ averaged over four higher resolution SAR subpixels:

$$\mathbf{K}_n = \sum_{s=1}^{4} \mathbf{K}_{ns}, \quad \mathbf{K}_{ns} = \mathbf{s}_{ns}\, \mathbf{s}_{ns}^{+} \tag{5.5-9}$$

Here, subpixels are indexed by two indexes, a pixel index n and a subpixel index s. The imaged areas contained simulated aircraft wrecks placed in forest and snow environments under foliage canopies.

The approaches to the NASA/JPL data study were dictated by the nature of the SAR data, in particular by its resolution and by the occurrence of phase miscalibration. The pixel size of the SAR imagery is 12.0 m in azimuth and 6.67 m in range. Since the wavelength of the radar is roughly 0.75 m, the resolution pixel spans many wavelengths. A large, man-made object, such as a metal plane or a corner reflector of several meters in size, can dominate scattering from a pixel. In many applications, large man-made structures produce a strong specular reflection, a large amplitude glint associated with the scattering of electromagnetic energy when the wavelength and pixel size are small compared to the characteristic length of the scatterer. These type specular returns are often used as a key for detecting targets. In the example considered, however, the area subsumed by a downed destroyed aircraft is considerably less than the area of a resolution cell so that the typical wreck of a small plane occupies only a fraction of a pixel. Due to the small size, most pieces of aircraft wreckage do not produce strong specular returns. As a result, "bright spots" in SAR imagery are not very useful for detection, resulting in a need to exploit polarimetric differences between the wreckage and the background.

Another complication in using the NASA/JPL SAR data is that it is not phase calibrated so the phase relationships between the scattered returns at the different polarizations are not consistent from image to image. The approach taken to circumvent the phase miscalibration problem is to perform adaptive target detection in each image independent of the other images. This also simplifies the problem in that we do not have to compensate for image-to-image variations arising from different depression angles, weather, and other measurement conditions.

This approach does result in one serious limitation, however, in that it is no longer possible to use the data from previously obtained images to train a neural network for application to new images. Since there are many clutter pixels on each image, the clutter model adaptation can be performed using a single image, but target models cannot be estimated, since there is only one target pixel on each image. In the approach taken the clutter is characterized independently on each image with the target being treated as an anomaly, being chosen as the pixel with the smallest likelihood of belonging to the clutter class.

Altogether, the NASA data set contained nine images containing one target each. Target positions were unknown. Clutter was strong relatively to signals, due to targets being under a heavy foliage and also due to the presence of ice in many of the images. The image size varied from $30 \times 50 = 1500$ pixels to $100 \times 100 = 10{,}000$ pixels. Two images from the NASA data set, the Michigan data and the North Carolina data, are considered in detail in Sections 5.5.3.1 and 5.5.3.2.

Another data set analyzed below was acquired using a Lincoln Laboratory 1-ft resolution polarimetric SAR. Records of the radar complex returns were available and the polarization measurements were calibrated. The data set was acquired near Stockbridge Massachusetts. We analyzed three strip maps ($3 \times 512 \times 2048$ pixels $= 1/6\,\text{km}^2$ at 0.75 ft per pixel sampling) of this data set. Available ground truth for this scene shows that within the range coverage of the SAR, there were nine military vehicle targets and nine corner reflectors (all but one of which were pointing within 90° of the SAR line of sight). The groundcover truth was available, indicating five types of groundcover clutter present in the scene: "treelines," "forest," "hedges," "roads," "fields," and the sixth type of clutter due to radar "shadows." Analyses of Stockbridge data are provided in Section 5.5.4. The discussion below begins with relatively simple models and evolves toward more complex models.

5.5.3 Physically Based Clutter and Target Models

5.5.3.1 Background Clutter Model

This section describes the background clutter model in a case in which there is no specific physical mechanism for a single or a few dominant scatterers in the pixel, so that a clutter pixel return is composed of multiple scatterers in that pixel, with random relative phases of the scatterers. Therefore, both I and Q components of the signal are sums of multiple positive and negative values, so that their distributions across similar types of pixels (same type of terrain) can be modeled as a Gaussian with zero mean. The phase of the return from the pixel, thus, may be considered to be uniformly distributed between zero and 2π, which is called a circularly symmetric distribution. Different types of terrain are described by Gaussian distributions with different covariances. Multiple types of terrain that might be present in an image are alternative sources of a signal, thus a probability distribution of signals in an image is modeled as a superposition of alternatives. We consider the complex scattering amplitudes to be statistically independent from pixel to pixel. This assumption is not necessarily valid, because nearby clutter pixels might contain contributions from the same scatterers. It is used here because the pixels subsume many clutter returns so that the assumption of clutter pixel independence may be used as an approximation.

According to these considerations, the clutter model assumes that the three complex scattering amplitudes $\mathbf{s}_n$ associated with a particular pixel n have a zero-mean, circularly symmetric, multivariate, complex Gaussian mixture density:

$$\text{pdf}\ (\mathbf{s}_n) = \sum_{m=1}^{M} r_m\, \text{pdf}\ (\mathbf{s}_n | m) \tag{5.5-10}$$

$$\text{pdf}\ (\mathbf{s}_n | m) = (\pi)^{-3}\ (\det\ \mathbf{C}_m)^{-1}\ \exp\left(-\mathbf{s}_n^{+}\ \mathbf{C}_m^{-1}\ \mathbf{s}_n\right) \tag{5.5-11}$$

where $\mathbf{s}_n^+$ is the Hermitian conjugate vector (this is complex conjugate and transposed), $\mathbf{C}_m$ are the complex covariances for the mode m, describing different types of terrain present in an image, and r_m are the priors or relative occurrences of the types of terrain. Adaptation of this model is achieved by estimation of the parameters r_m and $\mathbf{C}_m$. The circular symmetry property implies that the complex covariance matrices have a specific relationship between the real and imaginary parts of the elements, corresponding to the definition of these matrices:

$$\mathbf{C}_m = E\left\{\mathbf{s}_n \mathbf{s}_n^+\right\} \tag{5.5-12}$$

The exponent in expression (5.5-10) can be written as

$$\mathbf{s}_n^+ \, \mathbf{C}_m^{-1} \, \mathbf{s}_n = \mathrm{Tr}\left(\mathbf{s}_n \, \mathbf{s}_n^+ \, \mathbf{C}_m^{-1}\right) = \mathrm{Tr}\left(\mathbf{K}_n \, \mathbf{C}_m^{-1}\right) \tag{5.5-13}$$

According to Eqs. (5.5-10), (5.5-11), and (5.5-12), $\mathbf{K}_n$ is a sufficient statistic for the complex scattering amplitudes distributed according to the circularly symmetric complex Gaussian mixture distribution (5.5-10). An equivalent interpretation of this model is that $\mathbf{K}_n$ is distributed according to a mixture distribution of complex Wishart density functions:

$$\mathrm{pdf}\,(\mathbf{K}_n) = \sum_{m=1}^{M} r_m \, \mathrm{Wi}\,(\mathbf{K}_n | \mathbf{C}_m) \tag{5.5-14}$$

$$\mathrm{Wi}\,(\mathbf{K}_n | \mathbf{C}_m) = (\pi)^{-3} \, (\det \, \mathbf{C}_m)^{-1} \, \exp\left[-\mathrm{Tr}\left(\mathbf{K}_n \, \mathbf{C}_m^{-1}\right)\right] \tag{5.5-15}$$

In fact, in this example, data provided by JPL have been recorded as $\mathbf{K}_n$ averaged over four higher resolution SAR subpixels (5.5-9). Except for the lost resolution, this procedure is adequate, since several adjacent subpixels in most cases belong to the same type of clutter and are distributed according to a single mixture component (5.5-14).

The MLANS weights are computed as previously (5.2-10), with Wishart pdfs used instead of Gaussian ones. The estimation equations for rates (5.2-5) and (5.2-7) do not change. And the estimation (5.2-9) for covariances is replaced with

$$\mathbf{C}_m = \sum_n f(m|n) \, \mathbf{K}_n / N_m \tag{5.5-16}$$

5.5.3.2 Outlier Models

In some cases, relatively few Wishart components were sufficient to accurately model the clutter distribution, resulting in a few outliers of which one was the target. However, there were cases in which the outliers could not be unambiguously selected and additional processing had to be performed. Since these outliers are located in the tails of the likelihood distribution, an appropriate procedure must be brought to bear to obtain a more accurate representation of these tails. The application of Wishart mixture model provides the best representation of the central portion of the likelihood distribution because the majority of the observations are associated with that part of the distribution.

For the Wishart mixture model to provide a characterization of similar fidelity to the tails, a large number of components might be required. However, if the outlier observations could be separated from the remainder of the observations and the density function of the

outliers estimated in the absence of the central portion of the likelihood distribution then the distribution of the tails would be accomplished with a much smaller number of components. In removing the central portion of the likelihood distribution, it is apparent that although the total distribution may have zero mean, the outlier distribution will have nonzero mean, because outliers are displaced from the central portion of the zero-mean distribution. It is also possible that outliers are caused by large objects such as cliffs or by man-made objects such as cars or high power line towers. Thus, it is important that the likelihood function of the outliers be characterized by a mixture of components with nonzero means. Two models chosen for this purpose of a parsimonious representation of the tails of the distributions of the likelihood function are described in the next section.

Pixel Eigenvalue Model (PEM). The PEM model uses a multivariate Gaussian mixture density for the three eigenvalues $\lambda_n = (\lambda_{n,1}, \lambda_{n,2}, \lambda_{n,3})$ of the pixel covariance matrix, $\mathbf{K}_n$ (5.5-9). The pixel covariance matrix is Hermitian and its eigenvalues are realvalued and nonnegative; these considerations, on the one hand permit using real-valued mixture models given by Eqs. (5.2-7) through (5.2-9) and, on the other hand, indicate that Gaussian components may not be the best since they do not account for nonnegativeness. The PEM model does not convey all the information in the data because the eigenvalues provide only a portion of the available information in the data.

An important difference between the PEM model and Wishart mixture model is that the PEM components form a complete set of functions in the functional space of all possible pdfs of the eigenvalues, whereas Wishart components do not form a complete set of functions, because they are based on a zero-mean amplitude model (5.5-9). This issue of completeness is important if assumptions leading to the Wishart mixture are not exactly satisfied. Also, the PEM can be used in a supervised target detection approach, because the eigenvalues are unaffected by phase miscalibration.

Covariance Matrix Real Gaussian Mixture Model (CMM). The pixel covariance matrix $\mathbf{K}_n$, being a Hermitian matrix contains nine nonredundant real components: three diagonal components (K_{n11}, K_{n22}, K_{n33}), and real and imaginary parts of the three independent off-diagonal components (K_{n12}, K_{n13}, K_{n23}). The CMM model considers a pdf of a real-valued vector formed by these nine nonredundant components of the pixel covariance matrix $\mathbf{KR}_n = (K_{n11}, K_{n22}, K_{n33}, \mathrm{Re}K_{n12}, \mathrm{Re}K_{n13}, \mathrm{Re}K_{n23}, \mathrm{Im}K_{n12}, \mathrm{Im}K_{n13}, \mathrm{Im}K_{n23})$. The Covariance Matrix Real Gaussian Mixture (CMM) models the pdf of this vector as a multivariate Gaussian mixture. Parameters of this model are the rates, means, and covariances of the mixture components. The MLANS equations for this model are exactly the same as those considered above, Eqs. (5.2-7) through (5.2-9), with dimensionality $d = 9$, and $\mathbf{KR}_n$ used in place of λ_n. The components of CMM form a complete set of functions, this model does not utilize any specific a priori knowledge of scattering mechanism, and CMM utilizes all the information present in the pixel covariance data.

5.5.3.3 Rician Model

Models developed above are suitable for detecting a single target pixel in an SAR image as a least likely clutter pixel. If several target measurements are available, a two-class classifier can be developed that utilizes available information for each class. In case of a target present in an SAR pixel, the assumptions that led to the zero-mean circularly symmetric hypothesis for the clutter model may no longer be valid. Also, the Wishart mixture model is

not appropriate for clutter, if large clutter scatterers are present, such as cliffs or mountain ridges. A more appropriate distribution for target plus clutter is derived now following Marzetta (1995). In this case the radar pixel amplitude, $\mathbf{s}_n$, for the mth type scatterer can be written as a sum of the random clutter scatterers and a nonrandom one:

$$\mathbf{s}_n = \mathbf{x}_n + \mathbf{a}_m \exp(i\phi_n) \tag{5.5-17}$$

where $\mathbf{x}_n$, as in Eq. (5.5-9), is a circularly symmetric zero-mean complex Gaussian vector with covariance matrix $\mathbf{C}_m$, $\mathbf{a}_m$ is a complex deterministic but unknown three-dimensional scattering amplitude vector, and ϕ_n is the phase of this scatterer. The phase ϕ_n is related to the range from the radar to the pixel; it is very sensitive to the radar aircraft altitude and should be modeled as uniformly distributed over 0 to 2π and independent of $\mathbf{x}_n$. Therefore, the probability density of $\mathbf{s}_n$, conditioned on the mode m and on ϕ_n is

$$\text{pdf}(\mathbf{s}_n|\phi_n, m) = (\pi)^{-3} (\det \mathbf{C}_m)^{-1} \exp\left\{-\left[\mathbf{s}_n^+ - \mathbf{a}_m^+ e^{-i\phi_n}\right] \mathbf{C}_m^{-1} \left[s_n - \mathbf{a}_m e^{i\phi_n}\right]\right\} \tag{5.5-18}$$

Because the phase ϕ_n is random, the interest is in the probability density of $\mathbf{s}_n$ conditioned only on its mode membership

$$\begin{aligned} \text{pdf}(\mathbf{s}_n|m) &= \int \text{pdf}(\mathbf{s}_n|\phi_n, m)\, d\phi_n \\ &= (\pi)^{-3} (\det \mathbf{C}_m)^{-1} \exp\left\{-\mathbf{s}_n^+ \mathbf{C}_m^{-1} \mathbf{s}_n - \mathbf{a}_m^+ \mathbf{C}_m^{-1} \mathbf{a}_m\right\} \mathbf{I}_0\left(2 \cdot \left|\mathbf{a}_m^+ \mathbf{C}_m^{-1} \mathbf{s}_n\right|\right) \end{aligned} \tag{5.5-19}$$

where $\mathbf{I}_0$ is the zero-order modified Bessel function. It is seen that if $\mathbf{a}_m$ is phase shifted, the probability density is unchanged so that only the magnitudes and the relative phases of the components of $\mathbf{a}_m$ are important. A distribution (5.5-19) is called the Rice pdf. The parameters of the Rician mixture include rates, r_m, covariances $\mathbf{C}_m$, and scattering amplitudes $\mathbf{a}_m$. These parameters are estimated by the MLANS modeling subsystem. The ML neuronal estimation equations for the rates are the same as before:

$$r_m = \sum_{n=1}^{N} f(m|n)/N \tag{5.5-20}$$

The neuronal equations for the estimation of the covariance, $\mathbf{C}_m$, and scattering amplitude, $\mathbf{a}_m$, become

$$\mathbf{C}_m = \sum_{n=1}^{N} f(m|n) \left[\mathbf{s}_n^+ \mathbf{s}_n - \mathbf{a}_m^+ \mathbf{a}_m\right]/N \tag{5.5-21}$$

$$\mathbf{a}_m = \sum_{n=1}^{N} f(m|n) V_{nm} \mathbf{s}_n/N \tag{5.5-22}$$

Here, the weights $f(m|n)$ and V_{nm} are computed by the association subsystem. The weights $f(m|n)$ are computed as previously (5.2-10) using the Rice pdf (5.5-19), and V_{nm} are additional Rician weights that are the complex numbers with the amplitude

$$|V_{nm}| = \left[\mathbf{I}_1\left(2 \cdot \left|\mathbf{a}_m^+ \mathbf{C}_m^{-1} \mathbf{s}_n\right|\right) / \mathbf{I}_0\left(2 \cdot \left|\mathbf{a}_m^+ \mathbf{C}_m^{-1} \mathbf{s}_n\right|\right)\right] \tag{5.5-23}$$

and the phase factor

$$e^{i\phi_{mn}} = \left(\mathbf{a}_m^+ \, \mathbf{C}_m^{-1} \, \mathbf{s}_n\right)^* \Big/ \left|\mathbf{a}_m^+ \, \mathbf{C}_m^{-1} \, \mathbf{s}_n\right| \tag{5.5-24}$$

where $(\ldots)^*$ denotes the complex conjugate.

5.5.4 NASA Data Examples

The NASA data set was processed using the Wishart mixture model [(5.5-13) and (5.5-14)]. Rememeber that the NASA data set is comprised of nine images containing one target each with unknown target positions. All nine images were processed. In all cases the targets were identified without a single false alarm. In five cases the Wishart model processing results were deemed sufficient for target identification. In four other cases, we determined that additional processing was required using outlier models. Subsection 5.5.3.1 describes a typical example of the five cases using the Wishart model alone and how we determined that additional processing was not needed in this case. Additional processing for the other four cases is discussed in Subsection 5.5.3.2.

5.5.4.1 Michigan Data and Other Similar Cases

The upper row in Fig. 5.5-1 shows the results of applying the Wishart mixture model to SAR data taken in a heavily wooded area in Michigan. The first three images display the diagonal components of the pixel-covariance matrix data, $\mathbf{K}_n$ (these are the power in the two copolarized and in the cross-polarized signals). The likelihood function was computed from the entire complex pixel-covariance matrix data according to the Wishart model. The negative log-likelihood function image is shown next. The right-most image shows the thresholded negative log-likelihood function, which indicates that there are two outliers that may be considered as possible targets and in fact the most negative log-likelihood pixel was indeed the pixel containing the target. Thus, the target was detected without a single false alarm.

The number of mixture components (clutter types) was estimated as follows. The data were processed with four components. The estimated expected covariances $\mathbf{C}_m$ for each component were manually examined to detect any unexpected phenomenology. This was repeated using four, five, and six components. Nothing unusual was detected, but in the case of six components, two were very similar to each other. So it was decided to use five components in this example. A similar procedure was repeated for every image. From three to five components were required for different images. The process of the number of component selection can be easily automated.

Detection thresholds were selected as follows. A negative log-likelihood (NLL) histogram was plotted for each image (that is, the number of pixels vs. binned NLL values). Each histogram was manually examined. For five of nine images, the histograms could be divided into two nonoverlapping parts: most of pixels on the left (low NLL) and one to three pixels on the right. The detection threshold was selected to separate these few outlier pixels. For the other four images the histograms showed continuous distributions, where outliers could not be unambiguously detected. These four images were subjected to additional processing described in the next subsection.

5.5.4.2 North Carolina Data and Other Similar Cases

The first four images in the lower row of Fig. 5.5-1 show the North Carolina data set

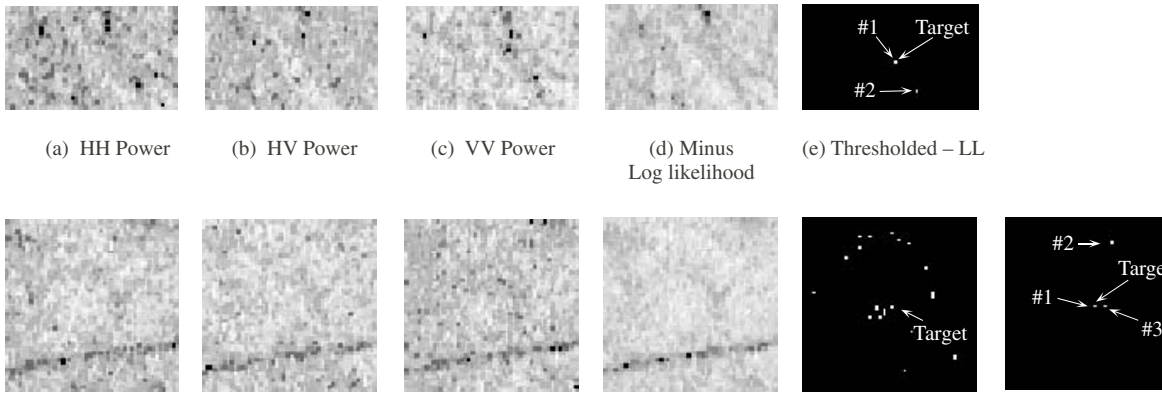


Figure 5.5-1 Identification of the downed aircraft wreckage under a heavy foliage canopy for two examples from the NASA data set. The upper row illustrates an application of the Wishart mixture model to the Michigan data setp; P-band SAR, 50×30 pixel image part is shown. The Wishart mixture model results in target detection without false alarms (**e**). The lower row illustrates the North Carolina data set; P-band SAR 50×50 pixel image part is shown. The first five images show the same type of data as the upper row. In this case, the Wishart mixture model does not lead to a reliable detection (**j**). The outlier model is used to reduce the potential number of false alarms, resulting in a successful target detection without false alarms (**k**).

processed using the Wishart model in a manner similar to the Michigan data set in the upper row. A threshold could not be selected to yield few outliers unambiguously. The fourth image illustrates results of this processing with a threshold being selected to yield 20 outliers. The PEM model with just a single component is applied to these outlier pixels. The right-most image illustrates the results: the characterization of the tails of the likelihood function distribution by estimating the distribution of the outlier pixels results in few "outliers of the outlier distribution." There were three obvious outliers, and the target is located in the pixel with the lowest likelihood. The target is detected without a single false alarm. Altogether, we have applied the outlier model processing to four data sets. In each case, we choose between the PEM or CMM models based on which one resulted in fewer remaining outliers. In every case, the least likely clutter point happened to be the target, so, all the targets were detected with zero false alarms.

5.5.5 Stockbridge Data Examples

5.5.5.1 Stockbridge Clutter Scene Segmentation

In the Stockbridge data case, clutter scene segmentation was of interest, as well as target detection, and a detailed characterization of the clutter types was performed using the Wishart mixture model. Wishart mixture model estimation automatically produces probabilistic image segmentation according to (5.2-10). (For the previous NASA data set, segmentation

is not discussed, because little is known about the image-truth clutter types.) An example of the Stockbridge scene segmentation is shown in Fig. 5.5-2. The minimum number of the required clutter-model mixture components was six, because there were six classes of clutter given by the image truth. Six components were determined to be sufficient to characterize clutter. This determination was based on comparing six-component segmentation results vs. image truth. An examination of the a posteriori probability plots for these components in Fig. 5.5-2 indicates that components 1 through 6 correspond to the various types of clutter present in the scene: "tree lines," "forest," "hedges," "roads," "fields," and "shadows." The components in this list are ordered according to the decreasing value of the estimated hh variance $[\sigma_{\mathrm{hh}}(m)]^2 = C_{33}(m)$ (5.5-16).

5.5.5.2 Stockbridge Target Detection

A much larger image area was processed in this example as compared to Section 5.5.3. Correspondingly, our approach to target detection in this case is more complicated than the one used there and is more representative of a real-time operation, when it is not feasible to process all data at once, but the data should be processed as they are acquired. A natural technique often used for this type of processing employs a sliding window. Within a sliding window, we estimate parameters of models for both the local clutter and target. Targets can be identified by thresholding the likelihood of the clutter mixture probability or by performing a likelihood ratio test. Whenever one implements a strategy that requires adapting to local clutter regions, some difficult issues must be addressed: (1) How is the clutter type estimated? and (2) How can we avoid biasing this estimate when the target is present but has not yet been detected? A conventional approach uses a sliding window with a guard region in the center (to avoid the target). This, however, creates a whole set of additional issues that must be addressed concerning the size of the window and the guard region (and their shapes and relative positions). Also, how, for example, are tree lines handled when the window overlaps two quite distinct regions of clutter? Then there are tradeoffs that must be made in choosing a window size that is large enough to collect meaningful statistics but small enough so that only one clutter type is present. It seems that more issues are brought up than resolved. Using MLANS permits us to avoid nearly all of these problems. A MLANS-based approach is compared to the traditional one in Fig. 5.5-3.

The recognition process uses a sliding window within which MLANS estimates *both* the clutter and target modes *simultaneously*. This solves the problem of biases that could be introduced by estimating only the clutter statistics when the target is present. Starting with initial estimates of the statistics of the potential clutter and target types, MLANS iteratively learns and adapts to provide a local estimate of the clutter types present and determines their relative proportions. Potential target pixels are identified as outliers by thresholding the likelihood of the clutter mixture probability. Alternatively, a two-class classifier can be invoked since MLANS has estimated the target statistics as well. In addition to providing the desired adaptation to clutter regions, our procedure also circumvents the problems attendant with using a guard region inasmuch as the target pixels are "captured" by the target modes of MLANS (which can be thought of as adaptive, complex-shaped guard regions).

We processed three strip maps of the Stockbridge data set containing $3 \times 512 \times 2048$ pixels. The sliding window size was chosen as $20 \times 20 = 400$ pixels, so that it covers a target plus sufficient number of pixels to estimate the clutter model. Available ground truth

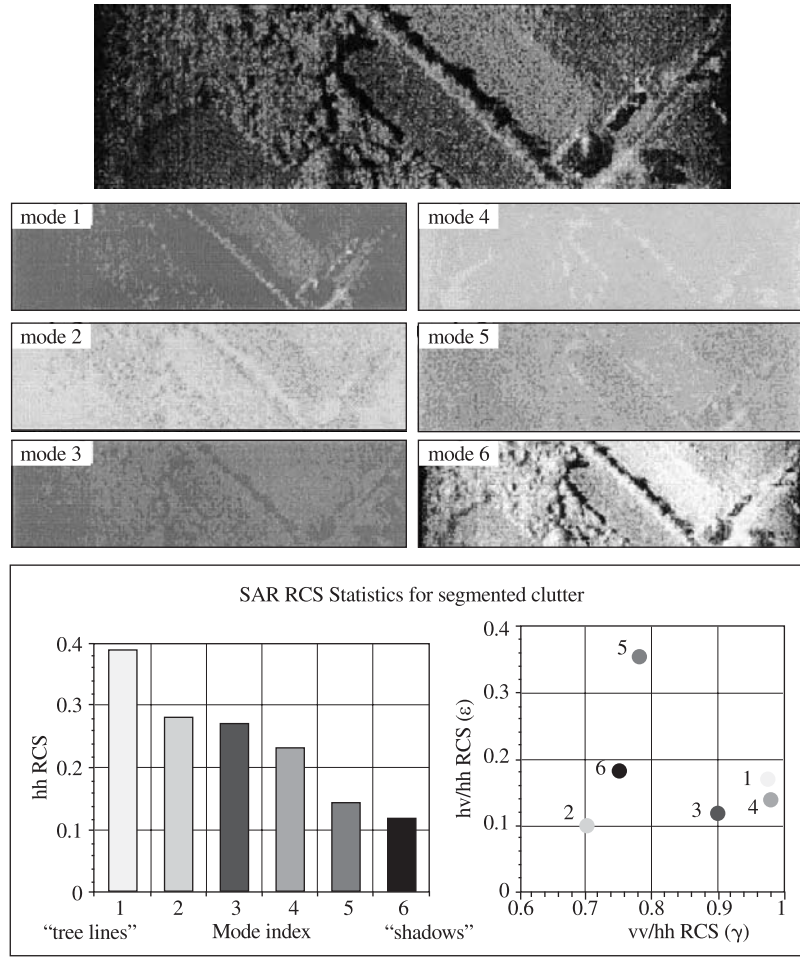


Figure 5.5-2 Wishart mixture model scene segmentation. The top image is the gray-scale composite of the a posteriori probabilities (shown below) for the six clutter modes that were automatically segmented by MLANS. The mean RCS values of the modes (bottom figures) show that the clutter requires a multimodal model.

for this scene shows that within the range coverage of the SAR, there were nine military vehicle targets and nine corner reflectors (all but one of which were pointing within 90° of the SAR line of sight).

MLANS detected seven of the nine targets (the two that were not detected were behind a tree line, hidden in shadows). In addition, MLANS detected two targets of which we had no previous knowledge and which were later confirmed to be actually present in the scene. All eight corner reflectors were also detected and identified as distinct from the targets. There was one not very credible false alarm (that is, just a few pixels were detected, while in each case of a true target a number of pixels were detected indicating a target-like shape). Detection was performed by thresholding the clutter likelihood, that is, by detecting the clutter-model outliers, similar to the discussion in Section 5.5.3. Using likelihood

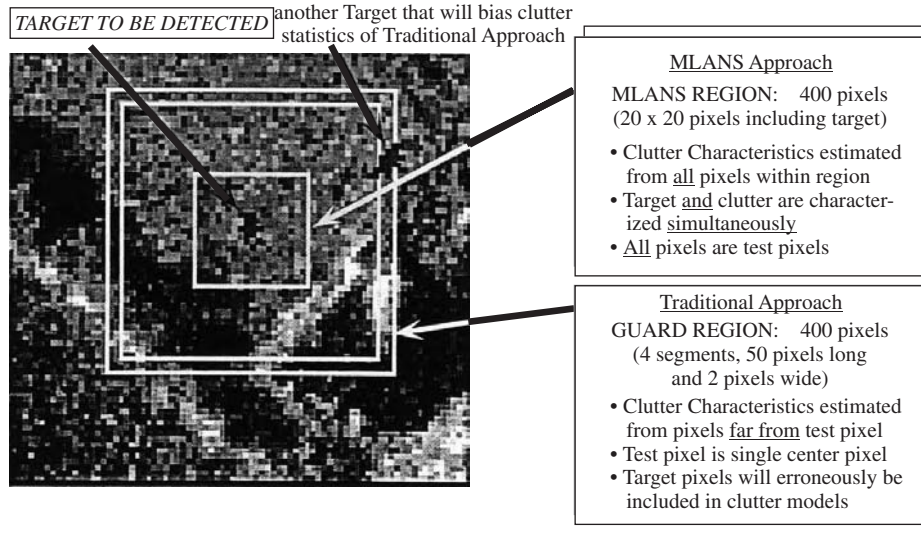

TARGET TO BE DETECTED
another Target that will bias clutter statistics of Traditional Approach
MLANS Approach
MLANS REGION: 400 pixels
(20 x 20 pixels including target)
• Clutter Characteristics estimated from all pixels within region
• Target and clutter are characterized simultaneously
• All pixels are test pixels
Traditional Approach
GUARD REGION: 400 pixels
(4 segments, 50 pixels long and 2 pixels wide)
• Clutter Characteristics estimated from pixels far from test pixel
• Test pixel is single center pixel
• Target pixels will erroneously be included in clutter models

Figure 5.5-3 Simultaneous estimation of the clutter and target distributions as a mixture model solves the problem of biasing class statistics. Clutter and target pixels are associated with appropriate classes, thus the presence of targets within the "sliding window" does not bias estimation of clutter statistics and vice versa. A much smaller window size can be used than in the "guard-region" approach, providing for better adaptation to the local scene characteristics.

ratio resulted in the same performance. Simultaneous with target detection, MLANS also segmented the scene into the clutter types that were consistent with the known ground truth.

The detection approach described in this section can also be used to collect data on targets in SAR images to facilitate the development of target models. As data are collected, the target model can be updated to improve the likelihood-ratio detection performance. The multipixel target data can then be related to physical scattering mechanisms to obtain a better understanding of the similarities and differences between target types, leading naturally to the development of multiple-pixel target models.

The development of multiple-pixel models could proceed in the same manner as the single-pixel models described above—with the exception that multiple pixels are used. That is, MLANS would operate in a higher dimensional space—the dimensionality being higher in proportion to the number of pixels being considered. In this way all pixel-to-pixel, polarization-to-polarization, and polarization-to-pixel correlations can be accounted for. This approach, however, leads very quickly to prohibitively high dimensional classification space. To avoid the "curse of dimensionality," a general model-based neural network approach described in Chapter 4 can be used. This will require the development of physically based multipixel models.

5.5.5.3 Stockbridge Data Analysis vs. Rician and Wishart Mixture Models

An assessment of the validity of the Rician model for man-made targets and a comparison with clutter models was performed using a subset of Stockbridge data. The results are shown in Fig. 5.5-4. The upper row of Fig. 5.5-4 illustrates clutter distributions estimated using

the Parzen method[1] (solid line), which in this case accurately represents the data, and using the two-component Wishart mixture models (dotted line). The first three plots (a, b, and c) show distributions of amplitude absolute values for three polarizations: $|s_{hh}|$, $|s_{vh}|$, $|s_{vv}|$ along the abscissa (Wishart components appear as Rayleigh in these axes). The last plot (d) shows similar distributions using Real(s_{hh}) abscissa (Wishart components appear as Gaussian). The natural clutter is well characterized by the Wishart model and requires two modes, one for the small cross section type clutter and one for the large cross section type clutter. The target data are shown in the lower row, plots (e, f, g, and h). The first three plots (e, f, and g) compare Parzen estimates of distributions for clutter and targets. The target data clearly show evidence of multimodal behavior that appears to be Rician rather than Wishart (as evidenced by the presence of peaks in the Parzen density estimates that are far from the origin). Some target data are shown in plot (h) using Real(s_{hh}) as an abscissa. A mixture of one Wishart component and three Rician components models this target data well.

5.5.6 Summary of SAR Models

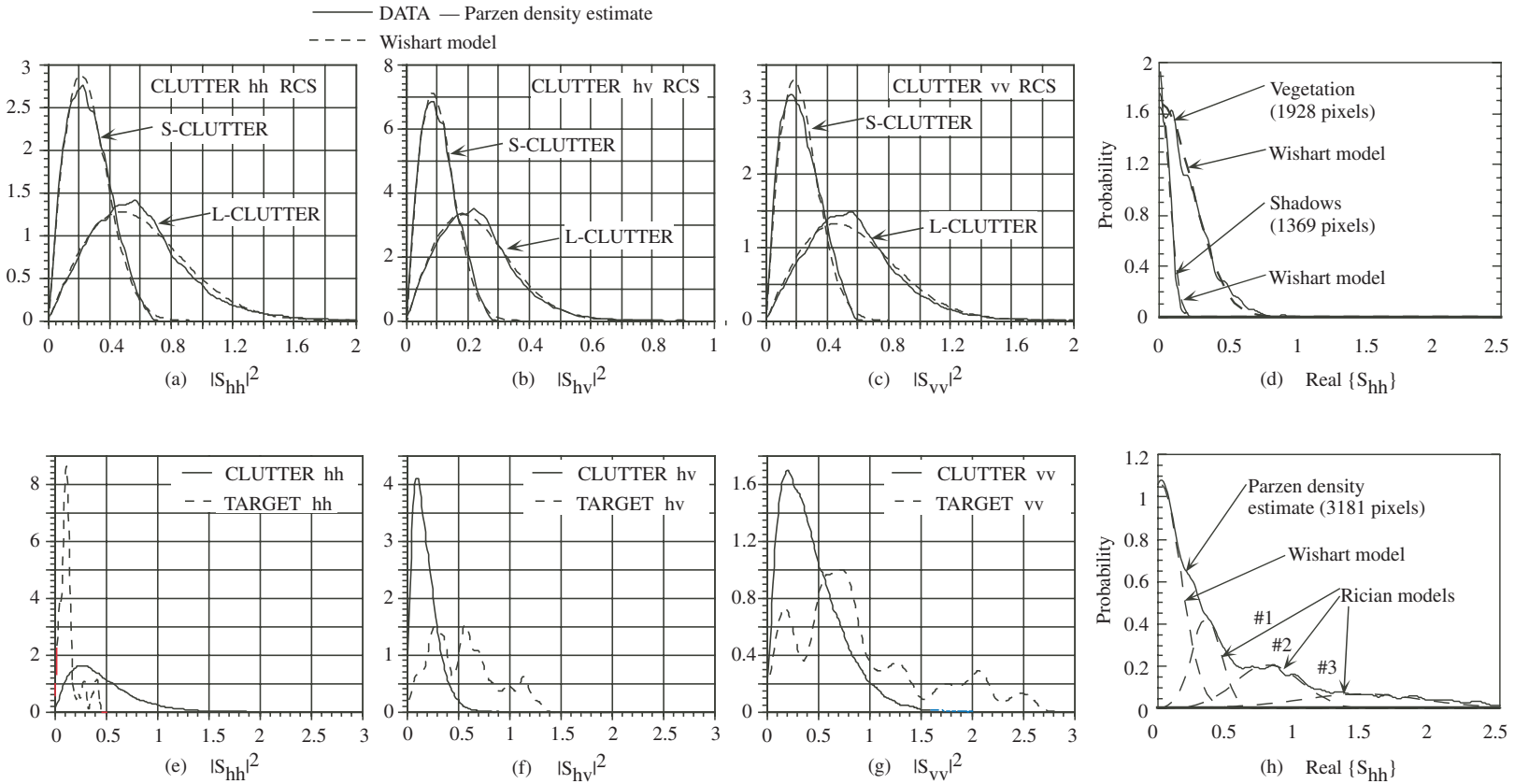
Although Gaussian models are widely applicable, they are not suitable for modeling radar signals because of their specific physical phenomenology. Several types of non-Gaussian mixtures have been derived in this section from the basic physical properties of SAR signals. These include complex Wishart and Rician mixture models that are appropriate for modeling individual pixel distributions in SAR data. More complicated models accounting for detailed deterministic information about objects, such as target shapes, can be developed using the general model-based neural network described in Chapter 4. More complicated models will be considered in the following chapters.

5.6 Convergence

In this section we derive the MLANS equations for the Gaussian mixture model and demonstrate that Eqs. (5.2-5) and (5.2-7) through (5.2-9) yield the ML estimation of the mixture parameters, while the MLANS weights are the a posteriori probabilities given by (5.2-10). These equations, of course, can be derived from the general model-based neural network equations in Chapter 4, Problem 5.2-4. Here, they are derived in a different way, providing a somewhat different angle on MLANS iterations. Then we consider convergence properties including the proof of the MLANS convergence.

5.6.1 Convergence and Learning

Convergence properties, discussed in this section, should not be mixed up with a related topic of learning abilities. Learning is an ability to extract maximum useful information from a given amount of data. More broadly, learning also includes an ability to find useful data. But what is useful information? Information has a precise mathematical definition discussed in Chapter 4; it is related to differentiating among alternatives. Differentiating more complicated choices requires more information. From this information-theoretic



TARGET PARAMETERS

MODE	PROB.	MEAN	STD. DEV
Wishart	33%	0	0.20
Rician #1	17%	0.40	0.15
Rician #2	19%	0.80	0.35
Rician #3	31%	1.60	0.65

Figure 5.5-4 Comparison of clutter and target data and their models. The upper row illustrates clutter distributions and Wishart models; (**a**, **b**, **c**) comparison of Wishart models (Rayleigh in these axes) and Parzen density estimates of the individual clutter components demonstrates the multicomponent Wishart nature of clutter (two components with a possible third component evident in the L-clutter distribution); (**d**) distributions of pixels classified by MLANS as vegetation and shadows; Wishart mixture model (Gaussian in this axis) fits data well. The lower row illustrates target data. Target and clutter distributions are compared (**e**, **f**, **g**); Parzen density estimates of the hh, hv and vv RCS magnitudes of the clutter and target single pixel data for the example scene clearly show the structural differences between the target and clutter distributions; it is clear that the target distributions cannot be modeled as Wishart mixtures, whereas Rician mixtures could be a plausible alternative; (**h**) target data distributions are plotted along with a mixture model utilizing one Wishart and three Rician modes.

view, "useful information" is a tautology. However, information as knowledge in general is not exhaustively characterized by mathematical Shannon information. For example, development of an accurate internal model of the world is useful, even though it may not be related to immediately important alternatives. But then, how do we know what is useful?—for example, by relying on a priori information. Thus, an accurate estimation of the parameters of a priori model can be viewed as extracting information from the world. We came to two types of ability to extract information: first, differentiating among alternatives and second, model parameter estimation. Theoretical tools for characterization of the information extracted for the parameter estimation are discussed in Chapter 9.

This section discusses a more preliminary and basic property of learning algorithms, convergence. A learning algorithm, given a finite amount of data, has to be able to come to an end of the learning process, or to converge. Is MLANS learning convergent? In Chapter 4, general MFT convergence was proved for incremental parameter changes during iterations ($\sim dt$). Here, we prove that Eqs. (5.2-5) through (5.2-10) are also convergent. How fast does it converge (in terms of the number of internal iterations or the number of elementary computational operations)? Does it always converge to the best possible solution or to a reasonable one? How often does it converge to a bad solution, under which circumstances, and how could this be mitigated? These questions were illustrated with examples throughout this chapter. Here we discuss some theoretical aspects of the MLANS convergence properties.

5.6.2 The ML Equations

The Maximum Likelihood (ML) estimation of parameters of Gaussian mixtures entails maximizing the likelihood L of all the observations in Eq. (5.2-6) over the set of parameters $\{r_k, \mathbf{M}_k, \mathbf{C}_k\}$. The derivation here is given only for an unsupervised case; because there is no formal difference between classes and modes, instead of two indexes k, m, only one index k is used for brevity. We maximize the loglikelihood $LL = \ln L$, which is equivalent to maximizing the likelihood L. A constraint on the priors r_k should be taken into account:

$$\sum_{k=1}^{K} r_k = 1 \tag{5.6-1}$$

This is a fundamental constraint, which says that the probability of all possible outcomes of any measurement $\mathbf{x}_n$ is 1. Constraint (5.6-1) can be accounted for by using a Lagrangian multiplier method. According to this method, the ML equations are derived by maximizing

$$LL' = \ln L + \lambda \left(\sum_{k=1}^{K} r_k - 1 \right) \tag{5.6-2}$$

over the set of parameters $\{r_k, \mathbf{M}_k, \mathbf{C}_k\}$ and a Lagrangian multiplier λ. This results in the following ML equations in addition to constraint (5.6-1):

$$\partial LL/\partial r_k + \lambda = 0; \quad \partial LL/\partial \mathbf{M}_k = 0; \quad \partial LL/\partial \mathbf{C}_k = 0 \tag{5.6-3}$$

Here, the unknowns are values of the parameters $\{r_k, \mathbf{M}_k, \mathbf{C}_k\}$ for a given set of data $\{\mathbf{x}_n\}$. Using the parametric expression for the likelihood in (5.2-2) and (5.2-6), the derivatives in (5.6-3) are calculated as follows:

$$\begin{aligned}
&\sum_{n=1}^{N} \left[r_k \,\text{pdf}\,(\mathbf{x}_n|k)\,/\text{pdf}\,(\mathbf{x}_n)\right] r_k^{-1} + \lambda = 0 \\
&\sum_{n=1}^{N} \left[r_k \,\text{pdf}\,(\mathbf{x}_n|k)\,/\text{pdf}\,(\mathbf{x}_n)\right] \left(\mathbf{C}_k^{-1}\,\mathbf{D}_{nk}\right) = 0 \\
&\sum_{n=1}^{N} \left[r_k \,\text{pdf}\,(\mathbf{x}_n|k)\,/\text{pdf}\,(\mathbf{x}_n)\right] \left\{\left(\mathbf{C}_k^{-1}\,\mathbf{D}_{nk}\right)^* \left(\mathbf{C}_k^{-1}\,\mathbf{D}_{nk}\right)\right. \\
&\left. -1/2\ \text{diag}\left[\left(\mathbf{C}_k^{-1}\,\mathbf{D}_{nk}\right)^* \left(\mathbf{C}_k^{-1}\,\mathbf{D}_{nk}\right)\right] - \mathbf{C}_k^{-1} + 1/2\ \text{diag}\left[\mathbf{C}_k^{-1}\right]\right\} = 0
\end{aligned} \tag{5.6-4}$$

In these equations, all vector and matrix indices are omitted for brevity: to avoid confusion all the vector-matrix dot products are included in parentheses, and * is used for outer products; in the last equation the derivatives with respect to the symmetrical matrix $\mathbf{C}_k$ are calculated according to Searle (1982).

The quantities

$$P(k|n) = \left[r_k \text{pdf}\,(\mathbf{x}_n|k)\,/\text{pdf}\,(\mathbf{x}_n)\right] \tag{5.6-5}$$

which appear in each of the above Eqs. (5.6-4) are the Bayesian probabilities of the nth observation to belong to the kth class, for every $n = 1, \ldots, N$, and $k = 1, \ldots, K$. In MLANS, these probabilities are association weights (5.2-10). The probability of an observation belonging to any of K classes equals 1 by the definition of $P(k|n)$ in Eq. (5.6-5):

$$\sum_{k=1}^{K} P(k|n) = 1, \quad n = 1, \ldots, N \tag{5.6-6}$$

It can be verified by direct substitution and using (5.2-10) and (5.6-6) that the MLANS Eqs. (5.2-5) and (5.2-7) through (5.2-9) are equivalent to Eqs. (5.6-4), satisfying constraint (5.6-1), with $\lambda = -N$, and keeping in mind that in this section k is equivalent to (k, m) in Section 5.2. But, it is also instructive to "derive" Eqs. (5.2-7), (5.2-8), and (5.2-9) from (5.6-4). For example, the first of Eqs. (5.6-4) can be rewritten as

$$\sum_{n=1}^{N} P(k|n) = \lambda\, r_k \tag{5.6-7}$$

Taking sums of each side of this equation over $k = 1, \ldots, K$, and using (5.6-1) and (5.6-6), we obtain $\lambda = N$, and (5.2-5). To obtain (5.2-6), let us rewrite the second of Eqs. (5.6-4) as

$$\sum_{n=1}^{N} P(k|n)\, \mathbf{C}_k^{-1}\, (\mathbf{x}_n - \mathbf{M}_k) = 0$$

After multiplying by matrix $\mathbf{C}_k$, this equation can be rewritten as

$$\sum_{n=1}^{N} P(k|n)\,\mathbf{M}_k = \sum_{n=1}^{N} P(k|n)\,\mathbf{x}_n$$

which is equivalent to (5.2-6). Equation (5.2-7) can be derived similarly, by noting that instead of the last of Eqs. (5.6-4) it is sufficient to consider a simpler equation,

$$\sum_{n=1}^{N} P(k|n)\ \left\{\left(\mathbf{C}_k^{-1}\,\mathbf{D}_{nk}\right)^{*}\left(\mathbf{C}_k^{-1}\,\mathbf{D}_{nk}\right) - \mathbf{C}_k^{-1}\right\} = 0$$

or

$$\sum_{n=1}^{N} P(k|n)\,\mathbf{C}_k = \sum_{n=1}^{N} P(k|n)\,\mathbf{D}_{nk}{}^{*}\,\mathbf{D}_{nk} \tag{5.6-8}$$

5.6.3 Local Convergence and EM Algorithm

MLANS unsupervised or partially supervised learning is an iterative process. Each iteration consists of the estimation of the parameters of object-type models followed by (or in parallel with) the computation of the association weights. Convergence means accomplishing learning in a final number of iterations. When parameters do not change much from iteration to iteration, this indicates convergence, as discussed in detail in Section 5.2.2.5. In our experience, MLANS usually quickly converges to a solution of the ML equations, within an order of 10 iterations after initiation and then one or two iterations thereafter, when pieces of new data become available. We prove now that MLANS learning is guaranteed to converge to a maximum of the estimated likelihood. This proof is related to the EM algorithm discussed in Chapter 4. The proof below uses Lemma 4.5.1, which was proved in Chapter 4. Recall,

LEMMA 4.5.1: Given $\sum_{i=1}^{I}\ q_i = 1, \quad \sum_{i=1}^{I}\ p_i = 1$. The $\max_{p_j}\left[\sum_{i=1}^{I}\ q_i \ln\left(p_i/q_i\right)\right]$ is attained at $p_i = q_i$.

We use the following abbreviated notations: $\mathrm{l}(\mathbf{x}|it) = \ln \text{ pdf } (\mathbf{x}|\mathbf{S}_{it})$, where $\mathbf{x}$ is a set of observations, $\mathbf{x}_n, n = 1, \ldots, N$; $\mathbf{S}_{it}$ is a set of model parameter estimates at MLANS iteration number it; $\mathrm{l}(n|it) = \text{pdf } (\mathbf{x}_n|\mathbf{S}_{it})$; $\mathrm{l}\,(k, n|it) = r_k \text{pdf } (\mathbf{x}_n|k, \mathbf{S}_{it})$; and we use $P(k|n, it)$ for a posteriori probabilities $P(k|n)$ computed at it-iteration. According to the definition of a posteriori probabilities (5.6-5), at every iteration,

$$\ln\ P(k|n, it) = \mathrm{l}(k, n|it) - \mathrm{l}(n|it) \quad \text{and} \quad \sum_{k=1}^{K} P(k|n, it) = 1 \tag{5.6-9}$$

Before proceeding with the proof of MLANS convergence, let us note that the estimated likelihood is finite. This follows from the fact that estimated covariance matrixes have to be invertible. (Such a constraint has to be imposed anyway for numerical reasons; see further

discussion in the next section.) To prove MLANS convergence, it is sufficient to show that the estimated likelihood or its logarithm increases from iteration to iteration. Because of the finiteness of the estimated likelihood, it follows that within a finite number of iterations, the estimated likelihood either stops changing or changes as little as desired (this is defined as convergence: the likelihood approaches a local maximum as close as desired). Now, examine a change of log likelihood between iterations:

$$\begin{aligned}
\mathrm{l}(\mathbf{x}|it) - \mathrm{l}(\mathbf{x}|it-1) &= \sum_{n=1}^{N} [\mathrm{l}(n|it) - l(n|it-1)] \\
&= \sum_{n=1}^{N} \left\{ \sum_{k=1}^{k} P(k|n, it-1) \right\} [\mathrm{l}(n|it) - \mathrm{l}(n|it-1)] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} P(k|n, it-1) \, [\mathrm{l}(k, n|it) - \ln P(k|n, it) - \mathrm{l}(k, n|it-1) \\
&\quad + \ln P(k|n, it-1)] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} P(k|n, it-1) \, [\ln P(k|n, it-1) - \ln P(k|n, it)] \\
&\quad + \sum_{n=1}^{N} \sum_{k=1}^{K} P(k|n, it-1) \, [\mathrm{l}(k, n|it) - \mathrm{l}(k, n|it-1)]
\end{aligned}$$

Here, in the second and third lines we used (5.6-9); the fifth line is nonnegative because of Lemma 4.5.1. The last line is also nonnegative, because of the following. Note that if $\mathbf{S}_{it} = \mathbf{S}_{it-1}$, the first item equals the second one so that this line is zero; but (5.2-7), (5.2-8), and (5.2-9) maximize the first item over $\mathbf{S}_{it}$, thus this line can only increase from zero up. This completes the proof of MLANS convergence at least to the local maximum of the likelihood function.

5.6.3.1 Monotone Likelihood Increase Example

The MLANS convergence and a monotone increase of the likelihood are illustrated in Fig. 5.6-1. In this example MLANS estimates parameters of a mixture of two one-dimensional Gaussian modes with parameters shown in Table 5.6-1. In Fig. 5.6-1 a contour plot (map) shows the $LL = \ln L$ values [Eq. (5.2-6)]. This plot was computed by using 200 data samples, $x_n n = 1, \ldots, 200$; 100 samples per mode were simulated according to the distributions in Table 5.6-1, and the LL value was computed for the model parameters values M_1 and M_2 varying between -200 and $+500$. The plot is symmetrical in M_1 and M_2 due to the symmetry of the Gaussian mixture in Table 5.6-1. The two maxima of LL correspond to the correct values of $M_1 = 200$ and $M_2 = 100$ and vice versa. The connected dots in this plot indicate the process of MLANS convergence from a difficult initial point along a valley right between the true distributions. We have started with a very bad, symmetrical initial guess, $M_1 = M_2 = 0$; because of the symmetry, after the first iteration, the estimated means are right at the center of the distribution: $M_1 \sim M_2 \sim 150$; then, in seven iterations, MLANS converges to the correct M_1 and M_2 values. In Fig. 5.6-2 the loglikelihood LL is plotted as a function of the iteration number, illustrating the monotone LL increase.

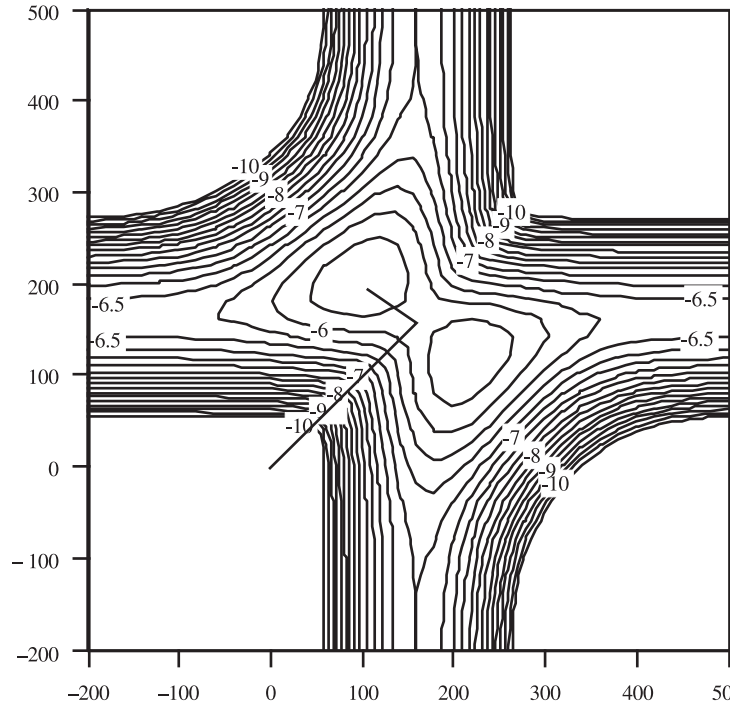


Figure 5.6-1 Illustration of MLANS convergence.

TABLE 5.6-1
Parameters of the Distributions in Figure 5.6-1

Class	Means	Rates
1	$M_1 = 100$	0.5
2	$M_2 = 200$	0.5

5.6.4 Global Convergence

A guaranteed convergence to a global maximum in a nonlinear estimation problem generally requires an exponentially large number of computations as a function of problem complexity. For example, a global search through all possible parameter combinations guarantees the convergence to a global maximum. More efficient search strategies have been designed, such as the thermal annealing algorithm. However, such computationally intensive procedures are not needed in practice. Within a hierarchical intelligence system, recognition is performed in a constant loop together with actions based on recognition results. Therefore, if at some moment an object or a set of objects is misidentified, either due to insufficient information or to convergence to a local maximum, the error will be found during the next action, and recognition results will be improved at the next recognition-action cycle. An example of such a MLANS operation was illustrated in Section 5.3.5.1.

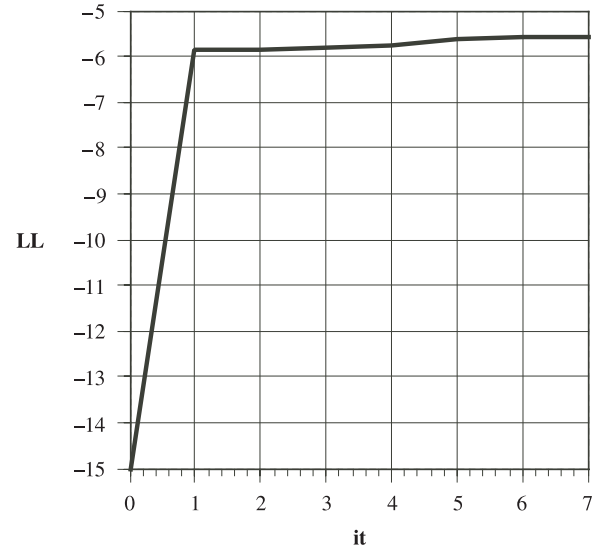


Figure 5.6-2 Illustration of the montone likelihood increase.

To improve finding the global maximum, two conditions should be verified on convergence. First, that there is no mode converging to a single data point, or, more generally, to a degenerate covariance matrix. A degenerate covariance corresponds to zero volume in the classification space, to infinite likelihood, and therefore to a maximum. A degenerate mode should be examined and either accepted as corresponding to a real object or rejected as a local maximum. A mode with nearly degenerate covariance has all its a posteriori probabilities either one or zero, and, therefore, these mode parameters do not change in subsequent iterations. Such a mode can be excluded from subsequent iterations to avoid numerical difficulties. Second, if two or more modes converge to the same mean and covariance values, they should be combined, their rates should be added, and MLANS should be let to reconverge.

5.7 MLANS, Physics, Biology, and Other Neural Networks

Let us look at MLANS as a simulation or model of a physical system. The basis for this is in a fundamental relationship between mathematical statistics and statistical physics. The log-likelihood function is analogous to $-H/T$, where H is the Helmholtz function (free energy) of the system and T is the temperature. The degrees of freedom identified with molecules in a physical system are identified here with parameters and neurons that estimate them. Some other types of neural networks have been identified with physical systems. For example, the dynamics of the Additive neural network (Grossberg, 1976; Hopfield, 1982) is similar to spin-glass systems. Spin-glass systems are governed by relatively simple Hamiltonians with local spin–spin interaction. Local interactions of spins correspond to the nearest neighbor classification computations in Additive neural networks. Complex

behavior in such systems is related to symmetry-breaking states. MLANS' more complex computational principles require analogies with correspondingly complex physical systems. Utilization of complicated a priori information in MLANS models corresponds to spatially noninvariant or nonsymmetrical Hamiltonians.

Temporal dynamics of the learning process in MLANS exhibits the existence of metastable states with long lifetimes, such as the initial large number of internal cycles shown in Fig. 5.2-9. From the physical point of view, this phenomenon is explained as a long relaxation-time process near a critical point of phase transitions. In MLANS, these phase transitions occur when the likelihood function has more than one maximum of approximately the same value. Phase transitions are related to establishing new concepts, and they are likely to occur when the number of objects available to the MLANS learning is small. In the case of nonstationary data, when new concepts should be constantly formed, MLANS can often be in a state close to a phase transition. Similar mechanisms have been postulated in the brain (Kryukov, 1988).

Such a situation is illustrated in Fig. 5.2-9, where the transitions between maxima evoke long-living metastable states. The learning process corresponding to Fig. 5.2-9 can be described psychologically as follows: as more information is acquired by the neural network, its initial internal representation of the "world" needs to be adjusted. This need for adjustment is an attentional mechanism that evokes short-term memory (STM, here is the metastable state), which is necessary for the modification of the long-term memory (LTM) containing the representation of the "world." Similar mechanisms for STM and LTM interactions exist in the ART neural network. This temporal dynamics of MLANS is intriguing from the biological point of view: would phase transitions in this neural network be helpful for understanding phenomena such as short-term memory and attention?

The nature of spatial symmetry and invariance violating Hamiltonians for modeling MLANS can be analyzed from the relationship between the geometry of a physical interaction and MLANS classification learning process. Interactions among molecules in a physical system are determined by their spatial proximity and correspond to inseparable terms in the Hamiltonian and Helmholtz functions. Analogously, parameters of classes, which are inseparable in the classification space, contribute to inseparable terms in the likelihood function corresponding to the overlapping distributions. Therefore, the closeness of the parameters of the classes and types of objects is determined by the overlap between classes in the classification space. MLANS weights form patterns of excitations in the classification space corresponding to the models. If classification models are combined with geometric models, as in image analysis, these patterns form dynamic spatiotemporal fields. This physical analogy can be used for designing physical devices implementing MLANS computations as a system of interacting fields. Particularly intriguing are quantum implementations of MLANS-like learning. Such quantum systems can be viewed as special-purpose quantum computers for pattern recognition.

The origin of the stochastic properties in MLANS is not in the stochastic properties of individual neurons, as the individual neurons do not possess any noisy quality in current implementations of MLANS. Stochastic properties of data are modeled explicitly in MLANS statistical models. In this respect MLANS is different from the Boltzmann machine neural network (Ackley et al., 1985). More similar to MLANS in this regard is the Mean Field Annealing neural network (Bilbro and Snyder, 1988). The role of temperature in MLANS is assumed by the covariance matrices; large a priori covariance matrices ensure initial

randomness or uniform probabilities of objects belonging to each class, analogous to a high initial temperature in an annealing process. Such a fuzzy initial state can be compared to Aristotelian Forms-as-potentialities, which are transformed into crisp concepts in the process of learning.

When comparisons are made to biological neural networks, it is necessary to consider if it is plausible that the brain performs complicated mathematical computations as does MLANS? The ML neurons perform complex computations such as matrix estimation and inversion, etc. These complicated ML neurons could be replaced by subnetworks of neurons each performing a simple operation. For example, a direct estimation of Choleski decompositions of inverse covariance matrixes can be obtained by a linear estimation procedure similar to the orthogonalization process, which is well suited for biological neural networks and which has been postulated by many researchers to take place in the brain (e.g., Grossberg, 1983). Thus, neurobiological analogies of MLANS ought to be sought on a higher, functional level, rather than on the level of individual neurons. For example, by estimating covariance matrixes of classes and types of objects, MLANS achieves an adaptive estimation of local metrics in classification spaces, allowing an adaptive enhancement of even minor differences between the objects, which are important for the classification, and an adaptive suppression of differences that are irrelevant for classification. It is clear that the brain performs similar functions. The following chapters consider even more sophisticated models built into the architecture of the model-based neural networks. These models along with estimation of covariance matrixes provide for a mathematical mechanism of adaptive metrics for matching models and data and for achieving subtle invariances that are problem dependent and adaptive in nature.

Learning mechanisms of neural networks include feedforward, feedback, cooperative, and competitive. The learning mechanism specified by Eqs. (5.2-5) through (5.2-10) is a feedback competitive learning. Feedback learning means that results of learning are used for future learning. Competitive learning means that neural weights "compete" with each other. Because MLANS weights are probabilities, class and type weights compete for probability of each observation. The architecture and temporal dynamics of MLANS can be compared with that of the Adaptive Resonance Theory (ART) neural network (Carpenter and Grossberg, 1987) in that MLANS is converging to a solution by "resonating" between input data and internal representations (models). This process accounts for the correlation of a current input with the neuron output at the previous cycle. Such correlations with delay have been considered for a long time as a more realistic replacement of the Hebbian learning rule. Adaptively building internal representations of the world based on a priori knowledge are a characteristic feature of both ART and MLANS. This combining of apriority and adaptivity was discussed in Chapters 2 and 4 as being important for establishing direct parallels between neural networks and philosophical concepts of mind.

MLANS models estimate spatially varying metrics in classification spaces, leading to classifiers of complicated shapes, which include multiple isolated regions. Isolated regions often correspond to different types of objects. The ability to learn the types on its own and to estimate a proper number of types for classification is referred to as the ability of a neural network to "conceptualize." MLANS learning can be compared with Aristotelian theory of Forms. According to Aristotle, Forms represent the a priori contents of mind, which have a potential for becoming concepts or categories of mind. This potential is realized in the learning process. MLANS implements a first step (albeit, in a simple way) toward

such interaction between a priori knowledge, learning, and concepts of mind: fuzzy a priori models become concepts or object-types in the process of learning.

MLANS is an efficiently learning network: as discussed in Section 5.7 and in Chapter 9, MLANS learning speed comes close to the theoretical limits of Cramer–Rao bounds for *any* learning algorithm. This results in a very good performance, close to the Bayes risk, with a much smaller number of training samples than is usually required by other algorithms or neural networks. An issue that is not resolved in MLANS models discussed in this chapter is recognition and intelligent processing of data in very-high-dimensional spaces, such as signals and images of resolved objects containing hundreds and thousands of samples, when the number of training samples is not sufficient for a reliable estimation of covariance matrices. This is related to the old unresolved problem of pattern recognition: how to extract classification features in an optimal way with insufficient information about class distributions. The general approach of a model-based neural network is to utilize more sophisticated a priori models combining adaptation with a priori information about physics, geometry, dynamics, and other properties of the objects. This development is continued in the following chapters.

NOTE

1. Parzen density estimation was briefly described in Chapter 1, Section 1.2. Recall, this is a nonparametric method of a nearest neighbor type, which does not require any assumption about data distribution. The Parzen method is suitable for data characterization in low-dimensional cases with sufficient data for accurate estimation.

BIBLIOGRAPHICAL NOTES

MLANS original development (Perlovsky, 1987b, 1988a; Perlovsky and McManus, 1991).

Statistical mixture models (Titterington et al., 1985).

Standard estimation techniques for Gaussian distributions (e.g., Fukunaga, 1972).

Inverse-Gaussian scale is also called Burdick plot (see Fukunaga, 1991).

Parzen pdf estimation; for further practical details see Fukunaga (1972).

Estimation of the number of agents, object-types, or clusters (see, for example, Anderberg, 1973; Fukunaga, 1988). For the ART neural network (Carpenter and Grossberg, 1987). For improving the internal representation of the world vs. making the best classification decision (Perlovsky and McManus, 1991).

Thermal annealing algorithm (Metropolis, 1980).

Number of clusters estimation (Yarman-Vural and Ataman, 1987; Fukunaga, 1988; Burdick and Perlovsky, 1991). Minimization of classification entropy (Perlovsky, 1987a; Perlovsky and McManus, 1991).

The content of Section 5.5 is based on the previously published work of a group of people (Perlovsky et al., 1997c; Marzetta, 1995; Schoendorf et al., 1994).

SAR operation, general (Goj, 1993); NASA SAR (Zebker et al., 1987).

Rice model (Rice, 1944, 1945).

MLANS vs. other neural networks: STM and LTM interactions in ART neural network (Carpenter and Grossberg, 1987); review of learning mechanisms of neural networks (Carpenter, 1989);

correlations with delay as a replacement of the Hebbian learning (see Klopf, 1987; Grossberg and Schmajuk, 1989).

Quantum implementations of MLANS (Chapter 8; Perlovsky, 1997c; Garvin and Perlovsky, 1995).

PROBLEMS

5.2–1 Write (5.2-1), (5.2-2), and the exponent in (5.2-6) in component notations similar to (5.2-4).

5.2–2 Write a computer code to plot a two-dimensional "$2-\sigma$ concentration ellipse," which is defined by $\mathbf{X}_n$ satisfying the equation

$$0.5\,\mathbf{D}_{nkm}^{\mathrm{T}}\mathbf{C}_{nkm}^{-1} = 2 \quad \text{(compare (5.2-6))}$$

Plot a few examples, e.g. (1) $\mathbf{M} = (0, 2), \mathbf{C} = \begin{bmatrix} 10 & 0 \\ 0 & 1 \end{bmatrix}$; (2) $\mathbf{M} = (0, 0), \mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$; etc.

5.2–3 Show that a Gaussian function is a δ-function in the limit of $\mathbf{C} \to 0$. Use the definition of a Gaussian function, the last row in (5.2-6), and a definition of δ-function, $\int \delta(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = f(0)$ for any (differentiable) function f.

5.2–4 Derive (5.2-7), (5.2-8), and (5.2-9) from general learning equations, Chapter 4 (take $dt = 1$). Compare to the derivation in Section 5.6.

5.2–5 Show that $\sum_{k;m} f(k, m|n) = 1$. [Use $f(k, m|n)$ definition (5.2-10).]

5.2–6 Consider a one-class, one-mode problem. Show that (5.2-7) and (5.2-8) are reduced to the classical expressions for estimating means and covariances (compare to Anderson, 1984). Interpret (5.2-9).

5.2–7 Write a computer code to compute the Bhattacharyya distance (5.2-11). Compute the Bhattacharyya distance for several pairs of distributions; see Problem 5.2-2.

5.2–8 Write a computer code to compute Gaussian distribution values according to (5.2-6). Validate the code by comparing with a calculator computation for few values of x in one and two dimensions; use a distribution from example 2 of Problem 5.2-2.

5.2–9 For two Gaussian distributions $G(\mathbf{x}|1)$ and $G(\mathbf{x}|2)$ with means $\mathbf{M}_1 = (0, 0)$, $\mathbf{M}_2 = (0, \sigma)$, and covariances $\mathbf{C}_1 = \mathbf{C}_2 = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$, compute likelihood ratios $LR_n = G(\mathbf{x}_n|1)/G(\mathbf{x}_n|2)$ for three data vectors, LR_1 for $\mathbf{x}_1 = (0, 0)$, LR_2 for $\mathbf{x}_2 = (-\sigma, 0)$, and LR_3 for $\mathbf{x}_3 = (0.5\sigma, 0)$. Use the code of the previous Problem 5.2–8.

5.2–10 Compute $P(1, 1|n)$ for $\mathbf{x}_n$, $n = 1, 2, 3$ from the previous Problem 5.2-9.

5.2–11 Write a code to compute OCs for two Gaussian classes of the previous problem: Step (1): simulate 100 samples $\mathbf{x}_n$, $n = 1, \ldots, 100$, for each of the two classes (use a standard routine to simulate Gaussian data). Step (2): for each $\mathbf{x}_n$ compute LR using the code of

Problem 5.2–9.; Step (3): sort *LR* values for each class (use a standard sorting routine). Step (4): vary threshold from min *LR* value to max *LR* value; for every threshold value, th, compute the number of class-1 data points classified as class-2 err1(th) and the number of class-2 data points classified as class-1 err2(th). Step (5): plot err1(th) vs. err2(th).

5.2–12 (Graduate level). Consider two Gaussian classes with distributions such that the first class is contained within the second class (in terms of their $2-\sigma$ concentration ellipses). Depending on the parameters of distributions (means, covariances, numbers of samples), the first class may form the largest cluster within the second class or may not. In other words, the ML clustering of the joint two-class distribution may result in an approximately correct solution: the smaller cluster to be close (within $1\,\sigma$) of the true mean of the first class, or may not. What is the probability of an approximately correct solution? See Fig. 5.2-5 and the related discussion. As a simpler problem, consider uniform distributions instead of Gaussian ones. Write a journal paper, for example, for Pattern Recognition Letters.

5.2–13 Verify the correspondence between Examples 2 and 3. Use (5.2-11), Table 5.2-1, Fig. 5.2-6, and an approximate correspondence of equal error rate of 2% and k-factor $= 3$.

5.3–1 Prove that (5.3-4) accomplishes the ML estimation in the case in which the teacher's information (P_{T}) is supplied as true probabilities independent from other data. *Hint:* consider modifications to likelihood (5.2-6) due to teacher's information.

5.3–2 Prove the first paragraph of Section 5.3.4. *Hint:* see Problem 5.3-1.

5.7–1 (Graduate level). Analyze a possibility of analog MLANS implementation by numerical simulation. Write a MLANS code and simulate a data set (e.g., use Problem 5.2-9 data set). Study MLANS performance for several levels of numerical accuracy (use declaration statements to modify numerical accuracy of all MLANS variables; explore REAL and INTEGER types for all MLANS variables). Identify "numerical accuracy bottlenecks" and develop workarounds, such as variable rescaling, local within each neuron, etc. Identify the most viable approaches. Write a journal paper, for example, for *Neural Network* journal.

5.7–2 (Graduate level). Study the effects of stochastic, noisy neurons on MLANS convergence in terms of the number of iterations and the attained maximal value of the likelihood. A first step here is to find a problem in which nonnoisy MLANS does not converge to the global maximum: consider an example of Problem 5.2-11; increase dimensionality; consider a more complicated example of Section 5.3.5.2. Consider an effect of mismatch between the data and models: simulate data using multimodal distributions. Consider complex real-world data. Write a journal paper, for example, for *Neural Network* journal.

EINSTEINIAN NEURAL NETWORK

This chapter describes the Einsteinian Neural Network (ENN) and its applications. ENN was introduced in Chapter 4. It is a model-based neural network, an instantiation of MFT. Its learning dynamics is determined by Shannon–Einsteinian similarity, which maximizes mutual information between the data and model. ENN models are inspired by the Einsteinian interpretation of the spectrum as a frequency density. This Einsteinian concept was extended in Chapter 4 to multidimensional spectra by considering them as densities of corresponding coordinates (frequency, time, space). Here this concept is applied to two-dimensional time–frequency spectra and images. We develop flexible compositional models of photon (or phonon) densities, suitable for solving inverse problems of identifying and estimating multiple overlapping signal sources.

We discuss ENN applications to classical spectrum estimation and to more complex transient phoneme-like spectra, to radar spectra, and to modeling ionospheric propagation of electromagnetic waves.

6.1 Images, Signals, and Spectra

6.1.1 Definitions, Notations, and Simple Signal Models

Usually, image is a two-dimensional spatial density of brightness or the number of photons, $S(x, y)$. Here, S is brightness or a number of photons and (x, y) are spatial coordinates. Color images can be characterized by one additional dimension, the frequency ω of each spatial pixel; thus, a color image is a three-dimensional density of the number of photons $S(x, y, \omega)$.[1] Signals are more simple objects than images; usually a signal is a one-dimensional time-dependent quantity, $s(t)$. An image brightness S is measured by an energy in a pixel, which is a positive number. Signal $s(t)$ could take positive and negative values. In signal analysis, it is often convenient to consider $s(t)$ as a complex number, characterized by two real quantities, amplitude a and phase ϕ,

$$s(t) = a(t)\exp[i\phi(t)]; \quad i = \sqrt{(-1)} \text{ is an imaginary unit} \tag{6.1-1}$$

The reason for this is that often $a(t)$ and $\phi(t)$ vary much simpler than signal $s(t)$: signal amplitude $a(t)$ varies slowly and $\phi(t)$ is a linear function, $\phi(t) = \phi_0 + \omega t$. In such a case, ω

is called an angular frequency (or just a frequency) of a signal $s(t)$; ω is measured in radians per second. For a pure-tone signal, $a(t) = \text{const}$. A signal representation given by Eq. (6.1-1) is convenient to use even if the measured signal is a single-component real quantity; in such cases complex numbers are used for convenience of analysis, and real signals are related to their complex counterparts through the relationship $\exp[i\phi(t)] = \cos[\phi(t)] + i\,\sin[\phi(t)]$,

$$s(t) = a(t)\text{Re}\exp[i\phi(t)] = a(t)\cos[\phi(t)] \tag{6.1-2}$$

In other cases, such as radar signals considered in Section 5.5, a measurement device measures two components of a signal, which are conveniently represented by a real and imaginary part of a complex signal. In the case of acoustic signals, $s(t)$ usually is given by pressure of the acoustic field, and an energy of a signal sample $S(t)$ is measured by (or proportional to) $|s(t)|^2$. Similar to light being composed of the quanta of electromagnetic field, photons, sound is composed of the quanta of acoustic field or phonons. In this book we will not be concerned with quantum properties of light or sound, but it will be convenient sometimes to refer to the numbers of photons or phonons. For acoustic signals, $S(t)$ is proportional to the number of phonons.

In a simple case of deterministic signals, $s(t)$ is described by a function, or by a differential equation with known deterministic parameters. An important class of signals is a superposition of several sinusoids, or more general, complex exponents

$$s(t) = \sum_k a_k \exp\left[i\,(\phi_k + \omega_k t)\right] \tag{6.1-3}$$

For example, simple musical sounds (say, of a single chord) can be well approximated by this type of model.

However, real world signals are often affected by noise and various other random effects. So, sometimes for practical reasons and sometimes because of fundamental properties of signal sources, signals are characterized statistically, by $\text{pdf}(\{s\})$. Here $\{s\}$ refers collectively to a set of continuous or sampled signal values over a time window. When parameters of this pdf do not change over time (from window to window), signal $s(t)$ is called stationary. An important example is sinusoids in white noise,

$$s(t) = \sum_k a_k \exp\left[i\,(\phi_k + \omega_k t)\right] + \text{wn}(t) \tag{6.1-4}$$

where $\text{wn}(t)$ is white noise. White noise is defined by statistically independent variabilities at different time points: $\text{pdf}\,[n\,(t_1)]$ and $\text{pdf}\,[n\,(t_2)]$ are independent if $t_1 \neq t_2$, in other words, $\text{pdf}\,[n\,(t_1), n\,(t_2)] = \text{pdf}\,[n\,(t_1)]\,\text{pdf}\,[n\,(t_2)]$. When parameters of these pdfs are independent of time, the signal is stationary.

Another important class of stationary signals that received wide consideration includes signals generated by a random noise input into a linear system. We briefly consider three types of signals from this class: autoregressive (AR), moving average (MA), and autoregressive moving average (ARMA). AR signals are given by a differential or difference equations

$$\sum_k \alpha_k\,(d/dt)^k s(t) = \text{wn}(t), \quad \text{for continuous case} \tag{6.1-5}$$

$$\sum_k a_k\,s\,(t_{n-k}) = \text{wn}\,(t_n)\,, \quad \text{for discrete case;} \quad k = 0, \ldots, K;\, \alpha_o = a_o = 1 \tag{6.1-6}$$

Here, α_k or a_k are constant coefficients of an autoregressive process, $s\,(t_{n-k})$ are sampled signal values, and K is called an order of the process. AR processes are also called Wiener filters, or Markov processes. They model deterministic linear systems, which parameters determine values of coefficients (α_k or a_k), driven by a random force, wn(t). AR processes can model oscillations, which are not as deterministic as sinusoids and exhibit random variations in periodicity. They have been used to model a large number of various systems and processes, including sunspot numbers and Brownian motion, and they continue to be used in modeling of economic trends, speech signals, and in many other fields.

Discrete moving-average (MA) signals are defined by

$$s(t_n) = \sum_k b_k \text{ wn}\,(t_{n-k}) \tag{6.1-7}$$

and ARMA signals are defined by combining AR and MA properties:

$$\sum_k a_k\, s\,(t_{n-k}) = \sum_l b_l \text{ wn}\,(t_{n-l})\,; \quad k = 0, \ldots, K; l = 0, \ldots, L; a_0 = b_0 = 1 \tag{6.1-8}$$

Stochastic stationary signals describe an important class of processes that was a subject of classical signal analysis. Virtually any book on signal analysis discusses properties of such signals and applicable estimation and modeling techniques in detail. Still, this is a restricted class: for many processes, including speech, statistical characteristics of signal pdfs change rapidly. Such signals are called nonstationary, or transient. Intensive development of mathematical techniques specifically for these more complicated processes is a relatively recent phenomenon. And ENN is developed for these more complicated nonstationary, transient processes. We will begin first with simpler, stationary signals, then extend our models toward nonstationary ones.

6.1.2 Frequency Components, Spectrum, and Spectral Models

Spectrum $S(\omega)$ is a density of signal energy in frequency. For stochastic signals, spectrum is given by an expected value of the square of the frequency component of the signal,[2]

$$S(\omega) = E\left\{|s(\omega)|^2\right\} \tag{6.1-9}$$

Periodicities in a signal are manifested as peaks in the spectrum. Spectrum estimation has been used for establishing periodicities since the nineteenth century, including periodicities in natural phenomena, economic indicators, etc. It is also important for analysis of various physical processes and in speech recognition. The most widely known and used method of spectrum analysis is Fourier transform (FT). Keep in mind that FT, in general, does not give a correct estimation of signals' frequency components (even though it is often thought that way). FT is a general nonparametric method that is suitable for signal analysis in nonstressing situations. However, it is not appropriate in stressing cases, for example, when one looks for separation of closely spaced frequency peaks, or for signals buried in noise or clutter, or for the analysis of nonstationary signals, whose frequency content (periodicities) quickly change over time. Examples of the latter include speech and financial market indicators. In stressing cases, parametric methods are more appropriate as they allow for utilization of a priori information about signal properties. Classical,

model-based methods are suitable for several relatively simple types of signals. The model-based spectrum estimation method discussed in this chapter utilizes flexible compositional models suitable for analysis of signals composed of multiple subunits, or multiple physical sources, overlapping in time and frequency (and also, if applicable, in space and propagation direction). These multiple signal sources could be the concepts that we attempt to recognize, such as objects in radar signals, or phonemes in speech recognition. A mathematical apparatus of modeling field theory for spectrum estimation is developed in this chapter using the mutual information similarity measure.

A most widely used method of spectrum estimation is based on Fourier transform (FT). FT represents a time signal $s(t)$ (see Note 2) as a weighted sum of $\sin(\omega t)$ and $\cos(\omega t)$ functions, or equivalently, by using complex exponents, $\exp(i\omega t)$,

$$s(t) = \sum_{\omega} s(\omega) \exp(i\omega t) \tag{6.1-10}$$

For sampled signals of finite length, time t and frequency ω span finite sets of values. Frequency spans a set of values determined by the sampling interval Δ and extent of time window T used for analysis. When a signal is sampled at time points given by

$$t = \{n \cdot \Delta, n = 1, \ldots, N\}, N = T/\Delta \tag{6.1-11}$$

frequency values are

$$\omega = \{n \cdot \Delta\omega, n = -N/2 + 1, \ldots, N/2\}; \quad \Delta\omega = 2\pi/T \tag{6.1-12}$$

when $T \to \infty$, ω spans the continuous spectrum. We would like to avoid using double indexes, so we would write $s(t)$ or $s(\omega)$ for $s\,(t_n)$ or $s\,(\omega_n)$.

An equation for computing FT from given signal values $s(t)$ is obtained from the definition (6.1-10), by using the following orthogonality property of the complex exponents (for the sets of t and ω values as given above):

$$\sum_{t} \exp[i(\omega - \omega')t] = N \cdot \delta_{\omega,\omega'} = (N \text{ for } \omega = \omega', 0 \text{ otherwise}) \tag{6.1-13}$$

Multiplying each side of Eq. (6.1-10) by $\exp(-i\omega' t)$, taking a sum over t, using the above equation, and changing $\omega' \to \omega$, we obtain

$$s(\omega) = (1/N) \sum_{t} s(t) \exp(-i\omega t) \tag{6.1-14}$$

A simple spectrum estimation technique, by taking a magnitude square of the FT, is called a periodogram,

$$\hat{S}(\omega) = |s(\omega)|^2 \tag{6.1-15}$$

This estimate is appropriate for a limited case of the signal $s(t)$ being a sum of sin or complex exp functions with frequencies exactly matching some of the set (6.1-12). In this case a periodogram will be nonzero only for the correct frequencies represented in the signal. In general, periodogram Eq. (6.1-15) has a number of limitations related to the properties of

stochastic signals as well as to the general limitations of nonparametric techniques. Let us briefly discuss some of these limitations. In a simple case of a signal being a single isolated sin (or exp) function with a frequency *not* matching any of the set (6.1-12), the periodogram will be nonzero for a range of frequency values around the frequency of the signal. When two or several closely spaced peaks are present in the true spectrum $S(\omega)$, they may not be resolved in the periodogram $\hat{S}(\omega)$. If there is a random component present in a signal $s(t)$, the periodogram does not converge to $S(\omega)$ even if the length of a signal goes to infinity.

Model-based spectrum estimation methods have been developed to surpass many of the limitations of the periodogram and many contemporary spectrum estimation methods are based on parametric spectrum models. Model-based methods enable utilization of a priori information about signal phenomenology and improve accuracy and resolution of spectrum estimation. Widely used models for spectrum estimation include spectral models for signals considered above: complex exponents in noise, AR, MA, and ARMA. Often, these models are successfully used for broad categories of signals deviating from these models.

To illustrate spectral properties of AR, MA, and ARMA signals, let us compute a theoretical spectrum of an ARMA signal. Substitute FT definition Eq. (6.1-10) into Eq. (6.1-8) defining an ARMA process:

$$\sum_k a_k \sum_\omega s(\omega) \exp(-i\omega t_{n-k}) = \sum_l b_l \sum_\omega \mathrm{wn}(\omega) \exp(-i\omega t_{n-l}) \quad (6.1\text{-}16)$$

By exchanging here orders of summation, using (6.1-10) and an orthogonality of an FT (6.1-13), the above equation can be rewritten as (see Problem 6.1-1 for details):

$$S(\omega) \sum_k a_k \exp(i\omega k \cdot \Delta) = \mathrm{wn}(\omega) \sum_l b_l \exp(i\omega l \cdot \Delta) \quad (6.1\text{-}17)$$

From this, FT of an ARMA signal, $s(\omega)$, can be expressed through a white noise FT, $\mathrm{wn}(\omega)$,

$$s(\omega) = \mathrm{wn}(\omega) \left[\sum_l b_l \exp(i\omega l \cdot \Delta)\right] \Big/ \left[\sum_k a_k \exp(i\omega k \cdot \Delta)\right] \quad (6.1\text{-}18)$$

The true (theoretical) value of a spectrum for white noise is constant, $E\left\{|\mathrm{wn}(\omega)|^2\right\} = \mathrm{WN}$ (this is the reason that the random uncorrelated noise is called white: like white light, it contains all frequencies with equal weighting). The true (theoretical) spectrum of an ARMA process is obtained by squaring the absolute value of Eq. (6.1-18) and taking an expected value,

$$S_{\mathrm{ARMA}}(\omega) = \mathrm{WN} \left|\sum_l b_l \exp(i\omega l \cdot \Delta)\right|^2 \Big/ \left|\sum_k a_k \exp(i\omega k \cdot \Delta)\right|^2 \quad (6.1\text{-}19)$$

This expression is often written by using z-variable

$$z = \exp(i\omega\Delta) \quad (6.1\text{-}20)$$

$$S_{\mathrm{ARMA}}(z) = \mathrm{WN} \left|\sum_l b_l z^l\right|^2 \Big/ \left|\sum_k a_k z^k\right|^2 \quad (6.1\text{-}21)$$

Thus, an ARMA spectrum is a ratio of two polynomials in z. A polynomial of Kth order has exactly K roots in a complex z-plane. Let us denote zeroes in the numerator as $z_{0,l}$ and in the denominator, as $z_{p,k}$. Equation (6.1-21) can be written as

$$S_{\text{ARMA}}(z) = \text{WN} \cdot \text{const} \cdot \left| \prod_l (z - z_{0,l}) \right|^2 \Big/ \left| \prod_k (z - z_{p,k}) \right|^2 \tag{6.1-22}$$

Let us briefly summarize properties of an ARMA spectrum. Zeroes in the denominator at $z = z_{p,k}$ are called poles of the spectrum; in general poles are complex values. If a pole $z_{p,k}$ value is real, the signal has a sinusoidal component; if a pole is inside the unit circle in complex z-plane, the signal has an exponentially damped sinusoidal component; if a pole is outside the unit circle, the signal has an exponentially growing sinusoidal component. If there are poles close to a unit circle, they define the gross characteristics of the signal and the spectrum. Since positions of poles are defined by an AR part of the signal model; the AR part is often a "more important" one. An MA part of the signal model is practically useful in providing extra degrees of freedom to the model: in a pure AR model, the ratio of amplitudes of various poles (residuals) is determined by positions of the poles. Thus, if an exact model of the signal is not known, AR could be too restrictive. AR and ARMA parameter estimation is practically important and theoretically intriguing, because there is no sufficient statistics for AR parameters and the Cramer–Rao bound for the accuracy of AR parameter estimation cannot be achieved for short-term windows. The meaning of this statement about the Cramer–Rao bound is discussed in detail in Chapter 9, but here it is sufficient to mention that *any* estimation of AR or ARMA parameters can be improved.

For nonstationary signals, spectra are changing with time, $S(t, \omega)$ and they are represented as two-dimensional images in (t, ω)-coordinates with brightness given by the spectrum S. A simple practical way to compute $S(t, \omega)$ is by subdividing signal $s(t)$ into short windows (t to $t + T$) and estimating the usual one-dimensional spectra for each window. The window length, T, should be selected so that signal spectrum does not change much over T. An elegant and general way of representing this process is given by the following expression for the (t, ω)-component of the signal:

$$s(t, \omega) = \sum_{t'} s(t') w(t - t') \exp(-i\omega t') \tag{6.1-23}$$

where $w(t - t')$ is a smooth window function that is concentrated within time interval $t' \sim (t \pm T/2)$ and is zero (or very small) outside of this interval. A periodogram-like estimate of the two-dimensional time–frequency spectrum is given by

$$S(t, \omega) = |s(t, \omega)|^2 \tag{6.1-24}$$

Images of the two-dimensional spectra $S(t, \omega)$ are also called spectrograms. They form the foundation for signal analysis in many areas dealing with nonstationary signals, for example, in speech recognition. Efficient modeling and estimation procedures for spectrograms did not exist until recently. This is the main subject of this chapter. In the following sections, we often consider data as given by $S(\omega)$ or $S(t, \omega)$, rather than by time signal $s(t)$. In some cases, measurement devices actually measure spectra[3]; in other cases, we assume that spectrum data are given by a procedure specified in Eqs. (6.1-23) and (6.1-24). Considering a

periodogram [rather than $s(t)$] as data does not change the conceptual nature of the spectrum estimation problem. Model-based spectrum estimation still consists in finding parameters of spectral models. These parameters could be used for finding specific events in spectral data, and spectral models could more accurately estimate the true spectra of signals.

6.1.3 Model-Based Spectrum Estimation

Model-based spectrum estimation consists in estimating parameters of spectral models. For an example of an ARMA process considered above, estimate parameters of Eq. (6.1-8) or (6.1-22): {WN, b_l, a_k}, or $\{\text{WN, const}, z_{0,l}, z_{p,k}\}$. The first model-based neural network for signal processing was Widrow's Adaline, which estimated parameters of an AR model. A number of parametric spectrum estimation methods have been developed, many of which are based on a fundamental statistical principle of the maximum likelihood (ML) for deriving parameter estimators. A maximum entropy spectrum estimation method was developed by Burg (1967). In this chapter we use an estimation approach based on Shannon's similarity, introduced in Chapter 4, where we discussed its relationships to other entropy maximization methods.

Our approach to the spectrum estimation as a $\text{pdf}(\omega)$ estimation is inspired by the Einsteinian interpretation of the electromagnetic spectrum as proportional to a pdf of the photon frequency. A similar interpretation is valid for phonons of acoustic spectra (speech, seismic signals, etc.) and for any signal field obeying Bose–Einstein's statistics (bosons). This chapter applies the Shannon–Einsteinian similarity developed in Chapter 4. The following sections introduce flexible models for $\text{pdf}(\omega)$ and $\text{pdf}(t, \omega)$.

6.2 Spectral Models

Einsteinian spectral models $F(\omega)$ were introduced in Chapter 4, Section 4.4. In general, they are superpositions of submodels, $F(\omega|k)$.

$$F(\omega) = \sum_k F(\omega|k), \quad k = 1, \ldots, K \tag{6.2-1}$$

Here, we consider several types of parametric expressions for the submodels $F(\omega|k)$. A most simple, uniform model is given simply by a normalization constant

$$F(\omega|k) = \hbar\omega\, A_k/N \tag{6.2-2}$$

A Gaussian model is given by Gaussian density $G(\omega)$. For continuous spectra, standard normalization is appropriate, $\int G(\omega)\, d\omega = 1$. For sampled spectra, models $F(\omega|k)$ are normalized appropriately, similar to Eqs. (4.4-3) and (4.4-10). The correspondence between the two normalizations is established by remembering that

$$\int d\omega(\cdot) \approx \sum_\omega \Delta\omega(\cdot); \quad \Delta\omega = 2\pi/T \tag{6.2-3}$$

Combining this with the Gaussian function (4.4-9),

$$F(\omega|k) = \hbar\omega\, A_k\, 2\pi/T\; G(\omega|k) = \hbar\omega\, A_k\, (2\pi)^{1/2}\, (1/T\sigma_k) \exp\left\{-0.5\; (\omega - \omega_k)^2\, /\sigma_k^2\right\} \tag{6.2-4}$$

In this equation, ω_k is the mean (center) frequency and σ_k is the frequency standard deviation for the source-model k. We will illustrate the use of Gaussian models in application examples discussed in Section 6.4, where these models are based on the physics of the considered spectra. Note that Gaussian functions form a complete set (in the space of positive functions), so that any spectrum $S(\omega)$ can be modeled as a superposition of Gaussian functions, using Eqs. (6.2-1) and (6.2-7).

We also consider spectral-pole models motivated by ARMA process. A prominent feature of these models is that their spectra contain poles in the complex frequency domain, determined by an AR part of the process. For a continuous-time process, a spectral-pole model in the complex frequency domain, corresponding to a first-order AR process, is a function of the type

$$\phi(\omega|k) = |\omega - \omega_k - i\alpha_k|^{-2} \tag{6.2-5}$$

Here, $(\omega_k \pm i\alpha_k)$ is a pair of spectral poles in the complex frequency plane, ω_k is their real part determining the frequency of an oscillatory process and a position of spectral peak along real frequency coordinate ω, and α_k is a damping factor, related to the width of the peak; vertical bars denote an absolute value (magnitude) of a complex number. To obtain an oscillatory process in a linear system with a real (not complex) coefficient, at least the second-order AR process has to be considered. It leads to a more complicated pole model containing two pairs of poles,

$$\phi(\omega|k) = \left|\omega^2 - i\alpha_k\omega - \omega_k^2\right|^{-2} \tag{6.2-6)/\$}$$

For a discrete-time process, sampled at the time interval Δ, a spectral pole model corresponding to the second-order AR process is a function of the type

$$\phi(\omega|k) = \Delta^4 \cdot |1 + a_k \exp(-i\omega\Delta) + b_k \exp(-2i\omega\Delta)|^{-2} \tag{6.2-7}$$

Parameters a_k, b_k of Eq. (6.2-7) can be related to parameters ω_k, α_k of Eq. (6.2-6) as follows (Perlovsky, 1988c):

$$a_k = -2\exp(-0.5\alpha_k\Delta)\, \cos\left[\left(\omega_k^2 - \alpha_k^2/4\right)^{1/2}\Delta\right], \quad b_k = \exp(-\alpha_k\Delta) \tag{6.2-8}$$

Using the above expressions, the pole models are defined as

$$F(\omega|k) = \hbar\omega\, A_k\, \text{const} \cdot \phi(\omega|k) \tag{6.2-9}$$

where *const* is determined so as to ensure normalization of the type (4.4-11): $\int \text{const} \cdot \phi(\omega|k)\, d\omega = 1$. It can be computed either numerically, or by using analytic expressions obtained in Perlovsky, (1988c):

first-order pole model, Eq. (6.2-5), $\quad \text{const} = 2\,\alpha_k/T$ (6.2-10)

second-order pole model, Eqs. (6.2-6) and (6.2-7), $\quad \text{const} = 2\,\omega_k^2\,\alpha_k/T$ (6.2-11)

For the discrete sampled process, the above expressions are approximations, valid for $\alpha_k \ll \omega_k \ll \Delta^{-1}$.

Gaussian and pole models given by Eqs. (6.2-8), (6.2-9), and (6.2-10), have a certain degree of similarity: when considered only over a set of positive frequencies $\omega = \{n \cdot 2\pi/T, n = 0, \ldots, N/2\}$, all functions have a single symmetrical peak, located at $\omega = \omega_k$, and all are characterized by two parameters, the mean position and the standard deviation or width of the peak. The second-order models are symmetrical about $\omega = 0$, but the first-order and Gaussian models are asymmetrical, indicating that they do not correspond to a linear system with real coefficients. Still, these models can be used as approximations for more complex systems.

6.3 Neural Dynamics of ENN

6.3.1 Shannon's Similarity Dynamics of Einsteinian Spectral Models

Spectral models introduced in the previous section are parametric models and their parameters are to be determined by maximizing Shannon–Einsteinian similarity between the models and data as described in Chapter 4. These models and the corresponding estimation equations define the Einsteinian Neural Network (ENN). The number of photons we denote N_ω. Its definition in Chapter 4 may present a difficulty at $\omega \to 0$, $N_\omega \to \infty$. This should be corrected as follows (see Note 3): in sampled spectra, frequency is defined within the sampling interval, $\pm\pi/T$. Thus, the smallest absolute value of ω is about π/T, and we substitute (4.4-2) with

$$N_\omega = S(\omega)/\hbar\omega'; \quad \omega' = \max(|\omega|, \pi/T) \tag{6.3-1}$$

Substituting Eq. (6.3-1) and any of Eqs. (6.2-4) or (6.2-12) into Eqs. (4.4-45) and (4.4-46), we obtain the dynamic equations of ENN for any of the above models. Fuzzy "class" memberships are given by

$$f(k|\omega) = F(\omega|k)/F(\omega); \quad F(\omega) = \sum_k F(\omega|k) \tag{6.3-2}$$

"Class" k here refers to a signal source submodel $F(\omega|k)$ of the overall model $F(\omega)$. And fuzzy class memberships $f(k|\omega)$ are interpreted as probabilities that a photon with frequency ω originates from source k. When deriving equations for amplitude parameters A_k, a constraint (4.4-12) should be accounted for. For all other parameters, $\mathbf{S}_k$,

$$d\mathbf{S}_k/dt = \sum_\omega N_\omega f(k|\omega)\ [\partial \ln F(\omega|k)/\partial \mathbf{S}_k] \tag{6.3-3}$$

In the case of Gaussian models, it is possible to derive more convenient iterative equations (4.4-13), (4.4-14), and (4.4-15). We also use approximations for the pole models described in the Appendix (Section 6.7), while here we summarize the results. Amplitude parameter estimation equations for all models are as follows:

$$A_k = N_k/N, \quad N_k = \sum_\omega N_\omega f(k|\omega), \quad N = \sum_\omega N_\omega \tag{6.3-4}$$

Here, N is the total number of observed (measured) photons, and N_k is interpreted as the number of photons coming from the source k. Equations for Gaussian model parameters ω_k and σ_k were derived in (4.4-14), and (4.4-15),

$$\omega_k = \sum_{\omega} N_\omega f(k|\omega)\omega / N_k \tag{6.3-5}$$

$$\sigma_k^2 = \sum_{\omega} N_\omega f(k|\omega)\ (\omega - \omega_k)^2 / N_k \tag{6.3-6}$$

Equations for the first-order pole model parameters ω_k and α_k are

$$\omega_k = \sum_{\omega} N_\omega f(k|\omega) \left[(\omega - \omega_k)^2 + \alpha_k^2\right]^{-1} \omega \Big/ \sum_{\omega} N_\omega f(k|\omega) \left[(\omega - \omega_k)^2 + \alpha_k^2\right]^{-1} \tag{6.3-7}$$

$$\alpha_k^2 = \sum_{\omega} N_\omega f(k|\omega) \Big/ \sum_{\omega} N_\omega f(k|\omega) \left[(\omega - \omega_k)^2 + \alpha_k^2\right]^{-1} \tag{6.3-8}$$

Equations for the second-order pole model parameters ω_k and αk are

$$\omega_k^2 = \sum_{\omega} N_\omega f(k|\omega)\omega^2 / N_k \tag{6.3-9}$$

$$\alpha_k = \pi \sum_{\omega > 0} N_\omega f(k|\omega)\ (\omega - \omega_k)^2 / N_k \tag{6.3-10}$$

The above Eqs. (6.3-2) through (6.3-10) define the dynamics of ENN with one exception: it is necessary to specify how many source models of each type $F(k|\omega)$ are included in $F(\omega)$, Eq. (6.2-8). This problem of determining numbers and types of model sources will be considered later. ENN consists of two subsystems, the association subsystem that computes fuzzy class memberships $f(k|\omega)$ according to Eq. (6.3-2) and the modeling subsystem that estimates model parameters according to Eqs. (6.3-3) through (6.3-10). ENN's architecture is similar to the general MFT architecture considered in Chapter 4 and to MLANS architecture considered in Chapter 5.

6.3.2 Two-Dimensional Time–Frequency ENN

Einstein's interpretation of the spectrum is extended in this section to the two-dimensional time–frequency spectra following Section (4.4–7) and applied to the time–frequency spectral model, $F(t, \omega)$. We interpret $F(t, \omega)$ as a pdf for a single photon with frequency ω at time t, which is proportional to a number of single-photon states at time–frequency (t, ω). A number of observed photons at each time and frequency $N_{t,\omega}$ is computed similarly to Eq. (6.3-1),

$$N_{t,\omega} = S(t, \omega)/\hbar\omega' \tag{6.3-11}$$

Let us specify several parametric models for signal sources. A uniform model, which could be appropriate for background noise, is given by a normalization constant,

$$F(\mathbf{x}|k) = \hbar\omega\, A_k / \mathrm{NCT}\Omega \tag{6.3-12}$$

where vector $\mathbf{x} = (t, \omega)$ and $\mathrm{NCT}\Omega$ is the number of cells (or pixels) in the $\mathbf{x}$-domain. For a rectangular (t, ω)-domain, $\mathrm{NCT}\Omega = \mathrm{NCT} \cdot \mathrm{NC}\Omega$, where NCT and $\mathrm{NC}\Omega$ are numbers of

cells (or pixels) along t and ω (in the one-dimensional case of the previous section we used simpler notations, $\mathrm{N} = \mathrm{NC}\Omega$). Nonrectangular domains can be described by $\mathrm{NC}\Omega(t)$. A single cell size in two-dimensions is $(\Delta\omega \cdot \Delta T)$.

A two-dimensional Gaussian model is appropriate for signal sources localized in time and frequency. It is given by (see Problem 6.3-1)

$$\begin{aligned} &F(\mathbf{x}|k) = \hbar\omega\, A_k\, (\Delta\omega)\, (\Delta T)\, G(\mathbf{x}|k) = \\ &\hbar\omega\, A_k\, (\Delta\omega)\, (\Delta T)\, (2\pi)^{-1}\, (\det\, \mathbf{C}_k)^{-1/2}\, \exp\left\{-0.5\, (\mathbf{x}-\mathbf{x}_k)^{\mathrm{T}}\, \mathbf{C}_k^{-1}\, (\mathbf{x}-\mathbf{x}_k)\right\} \end{aligned} \tag{6.3-13}$$

In this equation, $\mathbf{C}_k$ is the source model k frequency–time covariance matrix, and $\mathbf{x}_k$ is the source model mean vector, $\mathbf{x}_k = (t_k, \omega_k)$. Please note that the two-dimensional time sample ΔT here is a time duration between two short-term frequency spectra; if $F(t, \omega)$ is computed via short-term FT, each FT is typically computed over ΔT, so that $\Delta\omega = 2\pi/\Delta T$ and the original time sampling for the FT computation is $\Delta T/\mathrm{NC}\Omega$.

Four combinations of Gaussian and uniform shapes can be constructed in two dimensions of (t, ω): GU for Gaussian-time Uniform-frequency and, UG for Uniform-time Gaussian-frequency; the Gaussian parts of the models can be either with constant or varying parameters. For example, a varying-parameter UG model is given by

$$\begin{aligned} F(\mathbf{x}|k) = \hbar\omega\, A_k\, \mathrm{NCT}^{-1} (2\pi)^{-1/2}\, [\Delta\omega/\sigma_k(t)] \\ \exp\left\{-0.5\, [\omega - \omega_k(t)]^{\mathrm{T}}\, \sigma_k(t)^{-2}\, [\omega - \omega_k(t)]\right\} \end{aligned} \tag{6.3-14}$$

Such a model is appropriate for a source with localized frequency characteristics that vary with time. For certain sources, it is appropriate to consider combinations of Gaussian or uniform time models with pole frequency models.

Dynamic ENN equations are derived as in the previous section. Fuzzy class memberships are given by

$$f(k|\mathbf{x}) = F(\mathbf{x}|k)/F(\mathbf{x}) \tag{6.3-15}$$

They can be interpreted as probabilities that a photon at time t with frequency $\omega\, [\mathbf{x} = (t, \omega)]$ originates from source k. Equations for amplitude parameters A_k do not change; for all models they are estimated by

$$A_k = N_k/N, \quad N_k = \sum_{\mathbf{x}} N_{\mathbf{x}} f(k|\mathbf{x}), \quad N = \sum_{\mathbf{x}} N_{\mathbf{x}} \tag{6.3-16}$$

For all other parameters, $\mathbf{S}_k$,

$$d\mathbf{S}_k/dt = \sum_{\mathbf{x}} N_{\mathbf{x}} f(k|\mathbf{x})\ [\partial \ln F(\mathbf{x}|k)/\partial \mathbf{S}_k] \tag{6.3-17}$$

Again, for the considered models, it is possible to derive more convenient iterative equations. For example, two-dimensional Gaussian model estimation equations are given by

$$\mathbf{x}_k = \sum_{\mathbf{x}} N_{\mathbf{x}} f(k|\mathbf{x}) \mathbf{x}/N_k \tag{6.3-18}$$

$$\mathbf{C}_k = \sum_{\mathbf{x}} N_{\mathbf{x}} f(k|\mathbf{x})\ (\mathbf{x}-\mathbf{x}_k)\, (\mathbf{x}-\mathbf{x}_k)^{\mathrm{T}}/N_k \tag{6.3-19}$$

6.4 Applications to Acoustic Transient Signals and Speech Recognition

6.4.1 Transient Signals

Transient signals are encountered in many applications; a widely known one is speech recognition. Recognition of speech usually begins with recognition of phonemes, transient signals with duration about 0.1 to 1 s. A time-domain signal is sampled at a rate up to 20 kHz, producing thousands of samples per signal. A sampled signal is broken into windows of 10 to 20 ms duration, and for each window a fast Fourier transform (FFT) is computed, resulting in a (t, ω)-image of a speech sound. Then three problems should be solved: (1) determine the beginning and end of phonemes or other meaningful elements of speech (segmentation), (2) represent each phoneme with a small set of classification features (feature extraction), and (3) recognize phonemes. These three problems cannot be solved independently from each other; an iterative loop is required. Such an iterative loop often cannot be limited to a single phoneme: recognition of sequences of phonemes and words is used to refine phonemes. A comprehensive solution of speech recognition requires a hierarchical multilevel processing system with multiple loops of iterative processing: phoneme recognition is refined through recognition of words, and word recognition is refined through recognition of phrases and sentences, and so on. A vast literature is available on speech recognition. In this chapter, we concentrate on using ENN to improve solutions of the problems listed above: (1) segmentation and (2) feature extraction.

FFT of a signal is a set of complex amplitudes (alternatively, of pairs of real-valued sin and cos coefficients) for each frequency. Although some useful information is contained in the phase of the complex amplitudes, most of speech can be recognized from the absolute values or their squares. Such two-dimensional periodograms are often called Short-Term Spectra (STS).

Examples presented below illustrate the ENN concurrently solving problems (1) and (2). ENN segments STS and estimates parameters of the models. These parameters are used as classification features in subsequent processing. We first consider one-dimensional spectrum estimation for short signals, then we turn to two-dimensional spectra. At the end, we discuss a hierarchical architecture addressing problem (3), recognition.

6.4.2 Examples of One-Dimensional Spectrum Estimation

An important, fundamental issue in learning efficiency is how accurately model parameters are estimated from short-time windows. In the terminology of estimation theory, the question is: what is the efficiency of ENN as a parameter estimator? This question can be thoroughly studied when signals are generated according to a known model. The following examples demonstrate the efficiency and accuracy of ENN for AR parameter estimation and compare it with the performance of the best available classical estimators. AR model estimation is practically important and theoretically intriguing, because the Cramer–Rao bound for the accuracy of AR parameter estimation cannot be achieved for short-term windows and, thus, any existing estimator can be improved. In this well-studied area, where much progress has been achieved during more than 50 years of continuous development, the ENN efficiency

exceeds that of any other known estimator including the classical ones: the ML (Yule–Walker estimator, often known as the Levinson–Durbin algorithm) and the Maximum Entropy (ME) algorithm due to Burg.

To evaluate the ENN performance we have generated numerically a large number of AR signals of various orders with various parameter values. Typical results are shown in Fig. 6.4-1 presenting results of ENN spectrum estimation along with classical ML and ME techniques. The true signal is a sixth-order AR process with 10% additive noise. In this case ENN resolves all three spectral peaks, accurately estimating their positions and width, while classical techniques fail. It is known from experience that when using the classical ML or ME methods, it is necessary to utilize models of higher orders than the true order of the process. In the above example of a six-order process, the ML and ME methods resolved all three spectral peaks when 12 or higher order models have been used, while the ENN results have been accurate with three modes. Accurate estimation with fewer parameters is important for several reasons: first, it is more elegant, second, it requires fewer data, and third, it improves recognition when signal parameters are intended to be used as classification features, since it is desirable to reduce the number of features.

Similar spectrum estimation results are shown in Fig. 6.4-2 for the second-order AR process. In this case as well as in Figs. 6.4-3 and 6.4-4 a single-mode ENN model is compared with the second-order AR models estimated using classical techniques.

To systematically evaluate accuracy, efficiency, and robustness of ENN performance vs. classical techniques, we simulated thousands of signals for different values of modal parameters. Some of these results are shown in Fig. 6.4-3 summarizing results of 900 numerical examples evaluating the accuracy and precision of estimations of the frequency parameter of a second-order AR process [ω_k, Eq. (6.2-10) and (6.2-11)]. We used 100 independent realizations of the AR process of varying length from 16 to 4096 samples. In

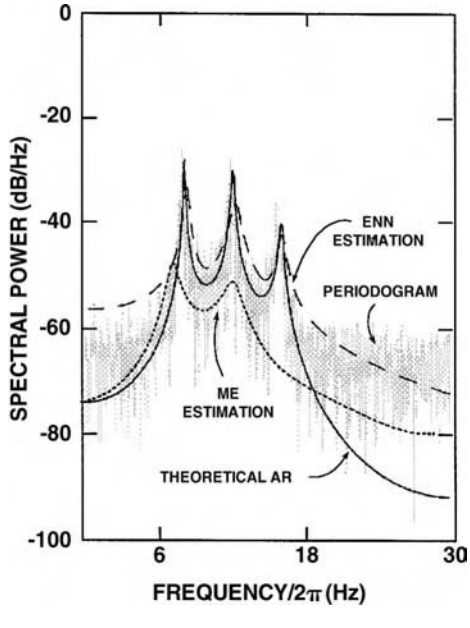


Figure 6.4-1 Spectrum estimation of a sixth-order AR signal. The known theoretical spectrum is shown as a solid line along with a periodogram (squared absolute value of a Fourier transform of a signal) shown as a shaded area. The ENN model utilizes three ARMA pole modes and sixth-order AR models are used for the ME and ML estimation. Results of the ENN estimation are clearly superior to the classical techniques (the ML estimation in this case was exactly same as the ME one).

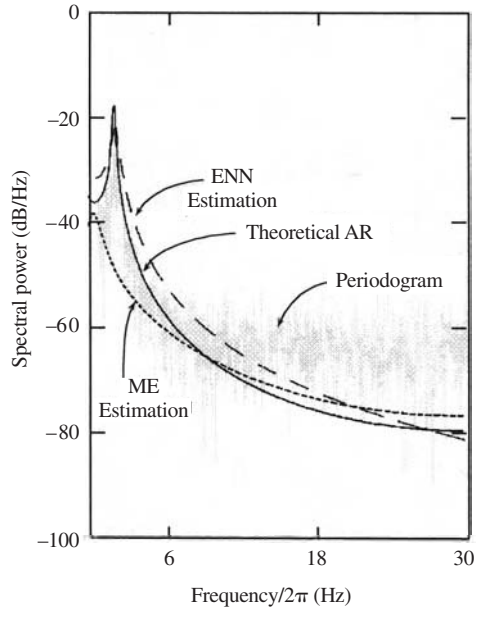


Figure 6.4-2 Spectrum estimation of a second-order AR signal is shown similar to Fig. 6.4-1. A unimodal ARMA model is used in ENN estimation and a second-order AR model is used for the ME and ML estimation.

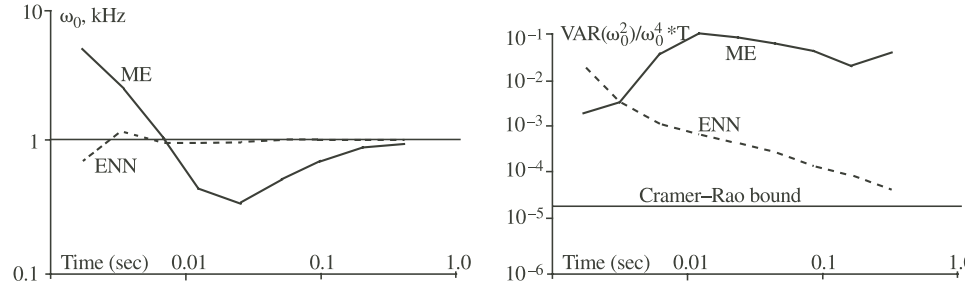


Figure 6.4-3 Numerical results of the estimation of the pole frequency ω_0 averaged over ensembles of 100 arrays. The true pole frequency $\omega_0 = 1$ kHz and the width of the pole (the imaginary part of the pole or a damping factor) $\alpha_0 = 10$ Hz; the sampling interval $\Delta = 10^{-4}$ s, and array lengths are from 16 to 4096 samples. A unimodal ARMA model is used in ENN and a second-order AR model is used for the ME and ML estimation. The ENN estimator is shown using dashed lines. It is clearly superior to the classical ML and ME estimators shown using solid lines.

this figure, the performance of the ENN is compared with that of the ME estimator for AR parameters. The Yule–Walker ML estimator in these cases yields the same results as the ME one. It is seen in Fig. 6.4-3a that the ENN estimator is more accurate than other available estimators. Its variance shown in Fig. 6.4-3b quickly tends to the Cramer–Rao bound, so that the MLANS performance is close to the bound on learning efficiency for any algorithm or neural network.

The ENN performance is also robust with respect to noise. When various types of noise have been added to the AR signals the ENN performance has been virtually unchanged,

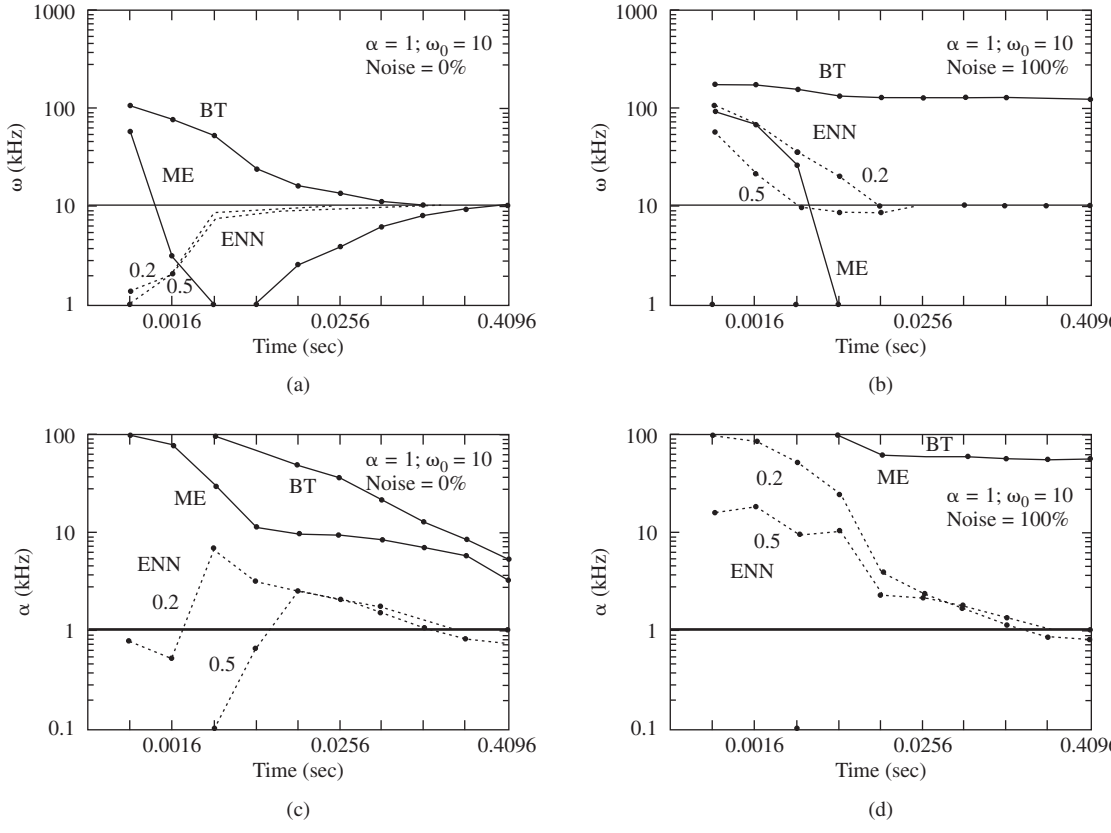


Figure 6.4-4 Numerical results of estimations of pole frequencies ω_0 and damping factors α_0 from AR signals corrupted with additive white noise. The true parameter values are: frequency $\omega_0 = 10$ kHz, damping factor $\alpha_0 = 1$ kHz, sampling interval $\Delta = 10^{-4}$ sec, and array lengths are from 8 to 4096 samples; (a) and (b) show results of frequency estimation, (c) and (d) show results of damping factor estimation; in (b) and (d) a white noise is added to the signal as described in the text. The ENN estimator is shown using dashed lines. It is clearly superior to the classical estimators shown using solid lines: the ML and ME estimators yield similar results in this case, BT denotes an estimator due to Bartlett (1946).

while classical ML and ME estimators have completely failed. We have studied additive and multiplicative types of noise with signal-to-noise ratios down to 1. Some of these results are shown in Fig. 6.4-4. The signal in Fig. 6.4-4a and c is a second-order AR process with the frequency $\omega_0 = 10$ kHz and the damping factor (the imaginary part of the spectral pole) $\alpha_0 = 1$ kHz. In Fig. 6.4-4b and d a white noise is added to the AR signal with the standard deviation equal to that of the AR signal (100% noise). Figure 6.4-4a and b shows estimation results for the frequency ω_0 and Fig. 6.4-4c and d shows estimation results for the damping factor α_0. In addition to the ME and ENN estimates, this figure shows the Bartlett (BT) estimates, which are modifications of the ML described in Bartlett (1946). Two curves shown for ENN correspond to different settings of an internal ENN threshold defined as follows. In Eqs. (6.3-9) and (6.3-10), N_ω is substituted with max $\{[N_\omega - \text{threshold}]\,, 0\}$; the *threshold*

was defined either as a minimum or average value of N_ω, which did not significantly affect the results. It is seen that the ENN performance significantly exceeds that of the classical estimation procedures: it is quickly converging and it is virtually unaffected by noise.

6.4.3 Two-Dimensional Time–Frequency Models

An example of a transient signal is shown in Fig. 6.4-5. This 1-s duration signal is sampled at 12.8 kHz rate and it contains 12,800 samples. Figure 6.4-6 shows the three-dimensional

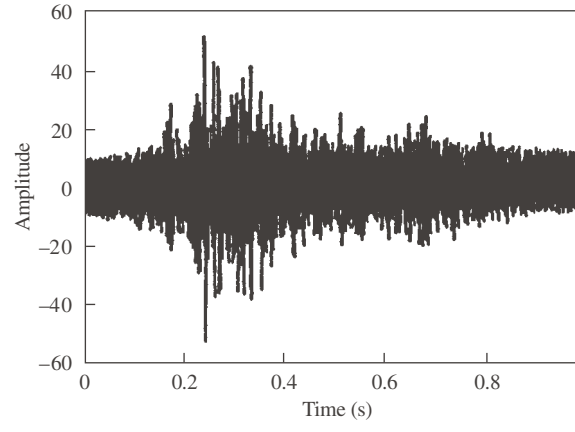


Figure 6.4-5 Example of a transient signal.

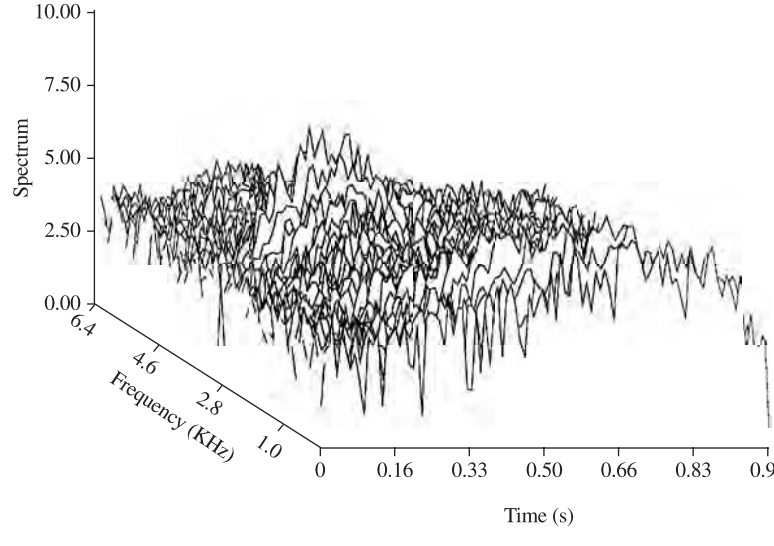


Figure 6.4-6 Short-term spectrum of the signal.

plot of the STS of this signal: a Fourier transform has been calculated for each 10-ms window and its amplitude squared is plotted along the vertical axis. The same data are shown as a contour plot in Fig. 6.4-7a. It is seen that there is a relatively short high-frequency component in the signal and a longer, low-frequency one with changing frequency. A random noise is scattered around the plot; the signal-to-noise ratio in this case is about one.

We modeled this STS data as a mixture of one uniform and several two-dimensional Gaussian models [Eq. (6.3-13)]. Results are illustrated in Fig. 6.4-7b and c, which shows estimated models with three and five Gaussian modes, respectively. Each Gaussian mode is illustrated by its $2-\sigma$ ellipse. (The number of ellipses in Fig. 6.4-7b and c is one less than the number of modes, since there is the uniform mode whose boundaries coincide with the plot boundaries.) Each Gaussian mode is characterized by six parameters: two components of the mean, two standard deviations, a correlation coefficient, and the energy in the mode. The energy in modes 4 and 5 is orders of magnitude smaller than in the first three modes, therefore the Fig. 6.4-7c model represents a small correction to the three-mode model in Fig. 6.4-7b in terms of signal energy. However, the importance of these modes for classification is not due to their share of the total energy, but to their effect on classification. Therefore, an optimal number of modes for classification can be determined only in a hierarchical system that consists of the signal-modeling layer considered here, and the top classification layer.

6.4.4 Hierarchical ENN + MLANS Architecture for Signal Recognition

A hierarchical system for signal recognition can be built from the bottom signal-modeling ENN layer considered above and the top recognition MLANS layer considered in Chapter 5. Such a hierarchical ENN + MLANS structure is shown in Fig. 6.4-8. The bottom ENN

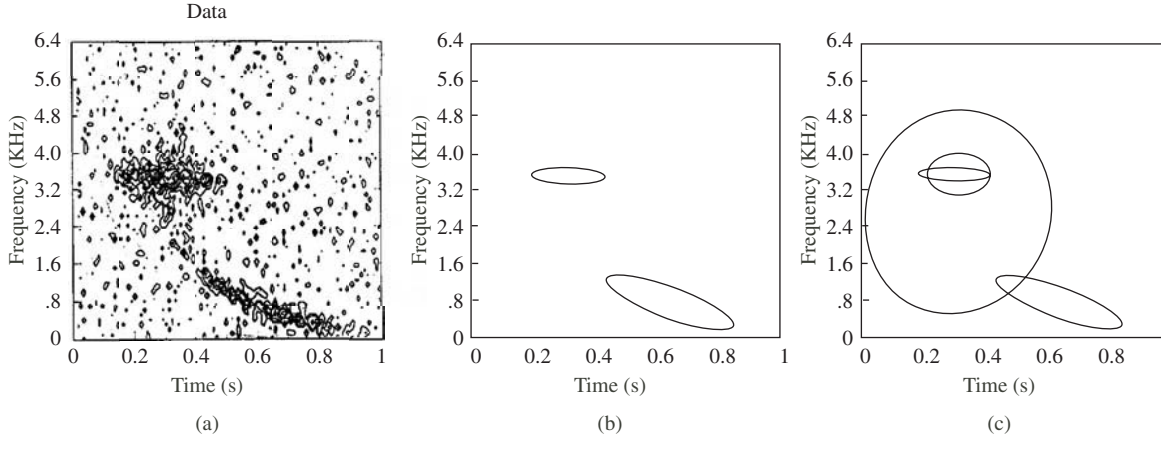


Figure 6.4-7 Transient signal model estimation. Contour plot of a short-term spectrum **(a)** and its modeling using three modes **(b)** and five modes **(c)**.

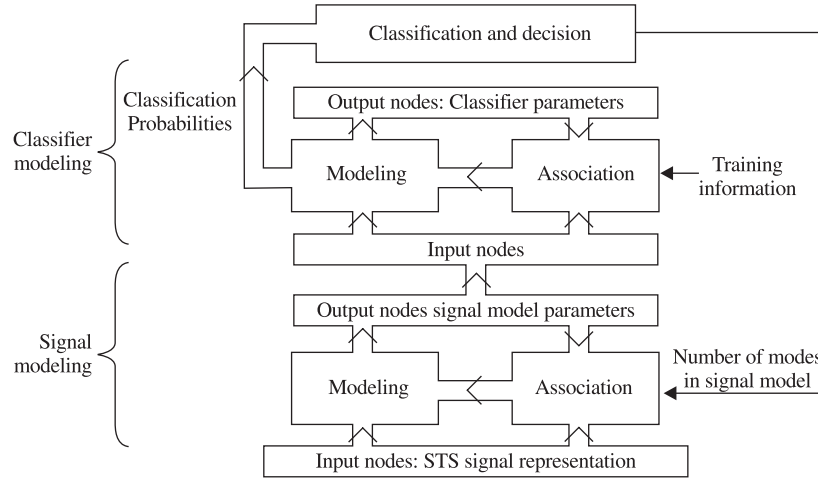


Figure 6.4-8 Schematics of a hierarchical ENN + MLANS architecture for transient signal modeling and classification.

layer estimates signal model parameters, which are used as classification features in the top MLANS layer. The top layer estimates feature pdfs and performs Bayes classification as discussed in Chapter 5.

6.5 Applications to Electromagnetic Wave Propagation in the Ionosphere

This section describes an application of ENN to characterizing a recently observed phenomenon known as equatorial ionospheric clutter that significantly affects propagation of high-frequency electromagnetic waves through the ionosphere and interferes with operations of over-the-horizon (OTH) radars and communication links using high-frequency radiowaves. Estimation of model parameters characterizing this phenomenon is complicated by the presence of multiple interfering signal sources. The data are affected by multipath propagation and scattering phenomena acting concurrently, each characterized by unknown parameters that vary in time and space. Therefore, the model of the data should be composed of multiple adaptive submodels, characterizing individual phenomena. This condition is not uncommon in science; data often are affected by multiple phenomena with unknown characteristics that cannot be isolated in scientific experiments. ENN resolves this complex estimation problem by probabilistic association of signal energy with submodels of various signal sources, while estimating model parameters.

6.5.1 Over-the-Horizon Radar Spectra

OTH radars operate in the high-frequency (HF) band between 5 and 30 MHz and are used to detect and track targets at distances up to 4000 km beyond the maximum line-of-sight

range of conventional ground-based microwave radars. To achieve these distances the HF signals propagate obliquely, reaching maximum altitudes ranging from 90 to 400 km in the ionosphere, and then reflect to the earth in what is called a bounce, or "hop." An illustration of OTH radar operations is shown in Fig. 6.5-1. Objects along the raypath scatter part of the radar signal energy back to the radar receive antenna. A "footprint" on the ground of a single radar resolution cell is approximately $10 \times 10\ \text{km}^2$. This is much larger than objects of interest, such as airplanes, thus most of the return signal is due to the ground clutter reflection. Relatively small returns from airplanes can be observed due to Doppler processing of the radar signal, which separates moving targets from the enormous ground clutter reflection. Doppler processing consists in computing short-term spectra: the ground return frequency is near the frequency of the transmitted signal (zero Doppler) while frequency of signals reflected by moving targets is shifted by Doppler frequency

$$\omega = (2v/c)\omega_0 \tag{6.5-1}$$

Here, v is the target velocity toward or away from radar, c is the speed of light, and ω_0 is the radar frequency. Equatorial clutter spreads over a number of Doppler cells and significantly degrades the OTH radar performance. For the basic studies of ionosphere, estimation of the parameters characterizing clutter is of interest.

Example of typical OTH radar Doppler spectra are shown in Fig. 6.5-2 for several time points, along with the processing results discussed later. Figure 6.5-3a shows same data for the entire data set containing 150 time spectra. The amplitude of returns is plotted using gray scale, as a function of Doppler frequency (horizontal axis) and time (or "spectrum number," vertical axis). Every horizontal line in this plot contains a 64-point Doppler spectrum.

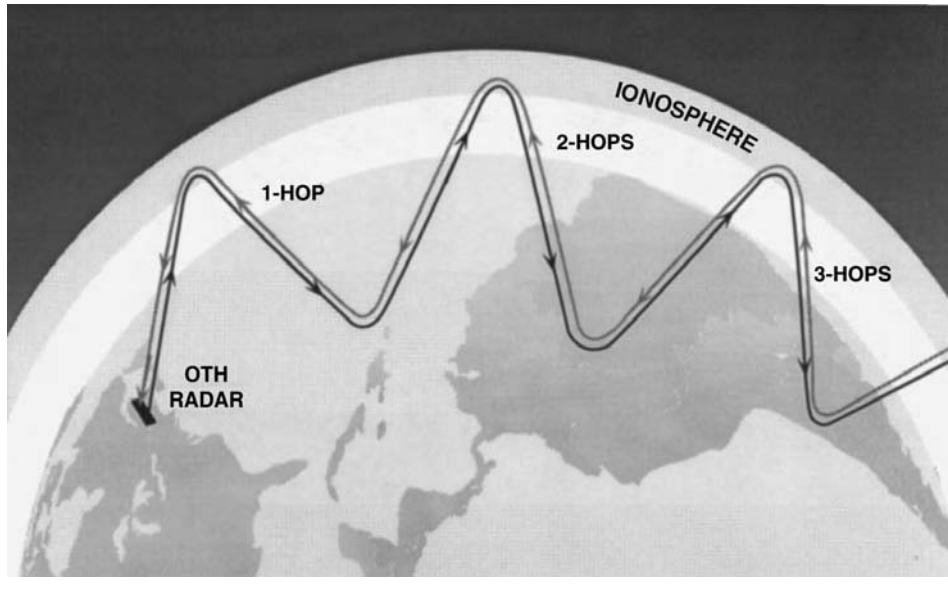


Figure 6.5-1 Artist's sketch of the OTH radar beam propagation with multiple hops.

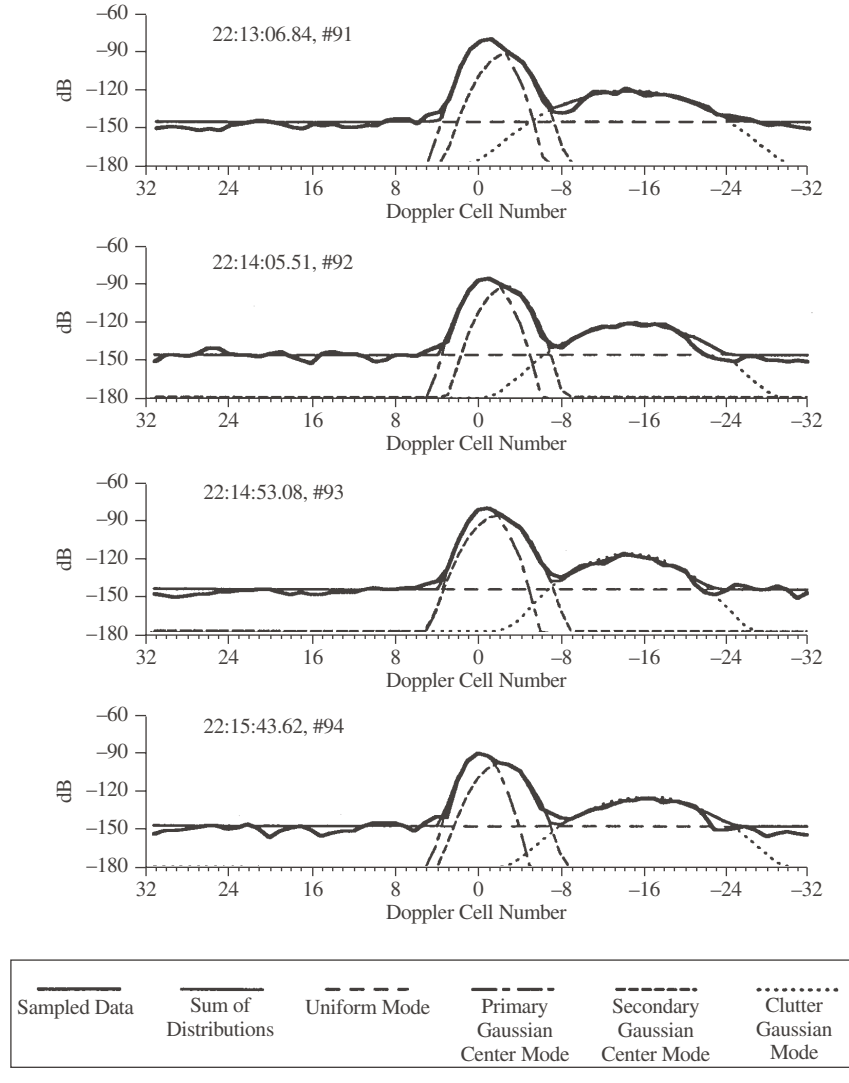


Figure 6.5-2 Comparison of estimated spectra vs. data for spectrum numbers 91–94, B34, RB1 data set. Individual modes are shown using dashed lines.

There are 150 spectra shown, numbered 0 to 149, all corresponding to the same range of 2856 km and collected from 21:07:44 UT to 23:04:58 UT. Each spectrum is determined by contributions from several major sources of signals. The main peak contains contributions first, from energy directly reflected back from the ground (or ocean) through the ionosphere to the receive antenna and second, from a "two-hop" mode that corresponds to a signal energy that bounces twice between the ground and the ionosphere (in one of the directions: away or toward the radar). The two-hop mode has lower energy and is slightly offset in

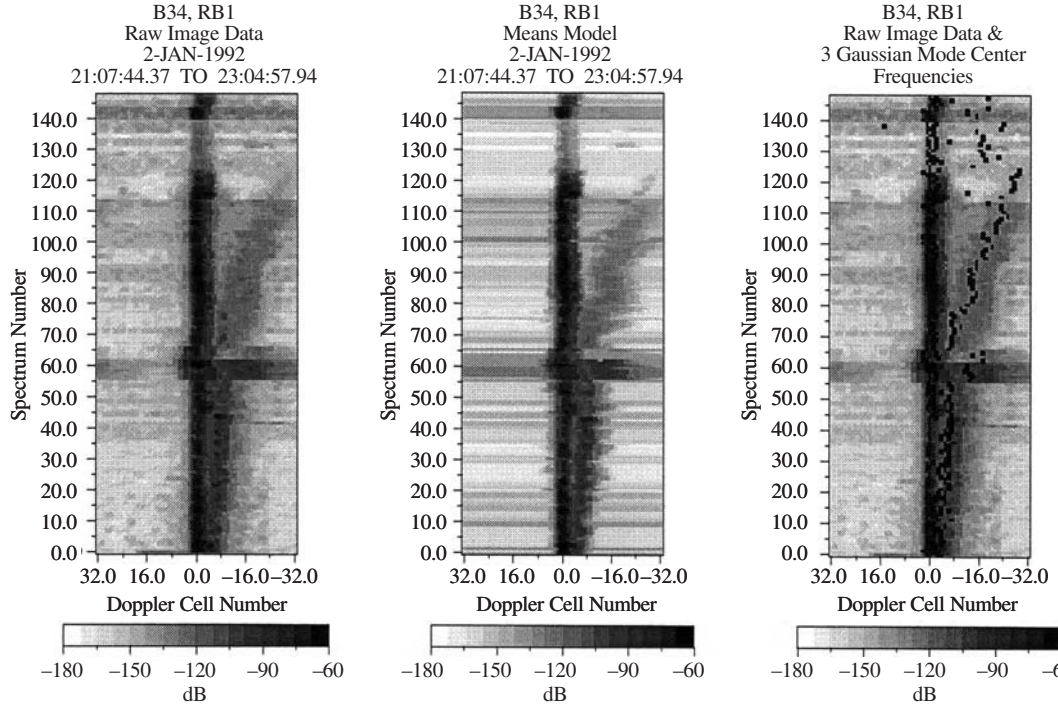


Figure 6.5-3 Comparison of estimated spectra vs. data for B34, RB1 data set; (**a**) data, (**b**) estimated ENN model, (**c**) mode-center frequencies superimposed on the raw data.

Doppler. [Similarly, there is signal energy that bounces several times (multihope modes) and has lower energy.] The vertical dark line corresponds to this ground return energy near zero Doppler; the Doppler is not exactly zero because of ionosphere motion. The spread-Doppler clutter structure, which is of main interest for our analysis, can be seen to the right of the ground return peak. This is caused by turbulent variations in the electron density in the region of the ionosphere from which the radar beam is being refracted. At time 21:07:44 UT the equatorial (also called spread-Doppler) clutter is barely discernible from the ground clutter and over the next one and one-half hours (toward the top of the figure, spectrum number 125, recorded at 22:42:26 UT) it evolves into a large structure spread over many Doppler values. (This severely impairs radar functioning; in fact, the radar operator actually changed the operating parameters of the radar for the time frames corresponding to spectrum numbers 58–65 and > 125 in an attempt to locate a better operating regime to improve the radar performance; however, this resulted in a significant loss of received energy.) The clutter also changes in range and azimuth (not shown), depending on the spatial variations in the ionosphere causing the spread-Doppler clutter. Another source of energy in these spectra is the noise "floor."

6.5.2 Spectral Models

For a long time, the origin of equatorial clutter in OTH radar spectra that spreads over significant Doppler interval remained a mystery. Early studies attributing these effects to the interaction of HF signals with either meteor trails or auroral irregularities could not account for the properties of clutter observed around the globe. Recently, Franchi and Tichovolsky (1989) obtained the Phase Screen Ground Modulation (PSGM) model, demonstrating how ionospheric turbulence can produce phase fluctuations in HF signals. They showed that the resulting effect can be modeled as a Gaussian-shaped spectral broadening, the parameters of which are related to a number of physically significant characteristics of electron distribution in ionosphere. Accurate estimates of the temporal and spatial characteristics of clutter spectra are needed for two purposes: first, for improved clutter rejection in normal radar and communication operations, and second to establish the quantitative connection between the observed clutter and ionospheric properties, for the basic scientific studies of the ionosphere. As mentioned, the necessary spectral clutter properties are difficult to estimate, because there are multiple sources of radar signals interfering with accurate estimation of properties of spread-Doppler clutter.

We model OTH Doppler spectra as a superposition of four major sources: (1) one-hop ground return, (2) two-hop ground return, (3) equatorial spread-Doppler clutter, and (4) noise floor. According to the PSGM model (Franchi and Tichovolsky, 1989), we use the Gaussian model Eq. (6.2-8) for the equatorial-clutter structure and for each ground-return mode. The noise floor is modeled using a uniform density [Eq. (6.2-10)].

Figure 6.5-2 shows the results of the ENN-estimated spectral model for these data. Four spectra and modeling results are shown for different time points. Estimated models for individual signal sources are shown using dotted lines. The total model is shown using a thin line and the data are in bold. There are almost no deviations between the data and the model. Figure 6.5-3 shows the same data and results for the entire RB1 data set containing 150 time spectra. Each one of the 150 horizontal lines in these plots is processed separately, so for every 64-point spectrum, ENN iterated until convergence, resulting in a set of model parameters (150 sets of parameters for 150 spectra). We illustrate the results by using these parameters to compute the estimated spectra models according to Eqs. (6.2-5), (6.2-7), and (6.2-8). The estimated models are shown in Fig. 6.5-3b. They are remarkably similar to the data, considering a relatively simplistic form of the model used. The model is seen to correspond to the overall data structure as it changes over time for both the ground-clutter and the spread-clutter structures. A comparison of the model parameters to the data is shown in Fig. 6.5-3c for the mean frequency parameters for each mode ω_k. The mean frequencies are seen to correspond to peaks in data for most of the 150 spectra for the ground-return clutter structure ($m = 2, 3$) and for the observed equatorial clutter event ($m = 4$). Model parameters can be related to ionospheric properties using the PSGM model and also can be used for developing of clutter rejection techniques as discussed in the next chapter.

6.6 Summary

This chapter described a new approach to spectrum modeling and estimation based on the Einsteinian interpretation of the spectrum as a pdf of frequency. The new estimation

principle is based on maximization of the Einsteinian likelihood. In Chapter 4 it was shown to be equivalent to the maximization of physical entropy of the ensemble of photons, and to maximization of mutual information in the model about the data. Also, in Chapter 4 we discussed the relationship of this estimation procedure to the classical maximum likelihood (ML) statistical estimation. The spectrum is modeled as a mixture of a set of basis functions. This estimation approach serves as a foundation for ENN, which is an MFT-type neural network based on Shannon–Einsteinian similarity.

The model-based neural network utilizes available prior information about signal properties. The ENN models are general and flexible; they are specific as little or as much as is warranted by available prior information. For example, Gaussian mixture models can be used to approximate *any* spectral shape. A practical utility of the model is high when relatively few submodels are sufficient for signal modeling. ENN learning is fast and efficient: for AR signal processing ENN estimation efficiency exceeds the best classical algorithms, the ME (Burg) and the ML (Yule–Walker) estimators even within the area of their applicability (AR signals). For two-dimensional time–frequency spectral modeling, ENN is efficient due to utilizing two-dimensional models with relatively few parameters.

An intriguing view of the ENN estimation process is that it corresponds to the quantum nature of the electromagnetic or acoustic field. Although signals considered in this chapter are due to acoustic and electromagnetic waves that can be considered as classical phenomena, the quantum nature of the fields determines statistical properties of a photon or phonon ensemble and the estimation procedure, in the same way as the Plank density is determined by the quantum structure of blackbody radiation.

6.7 Appendix

Shannon–Einsteinian similarity or mutual information [Eq. (4.4-44) is exactly similar to the Bayesian similarity (4.3–10), with the sum over pixels $\Sigma_n(\cdot)$ being replaced by the sum over photons $\Sigma_\omega\ N_\omega(\cdot)$. The Einsteinian models in this chapter can be considered as likelihood models for the frequency, ω, while every photon is considered as an independent observation of frequency. Thus, maximization of Shannon–Einsteinian similarity for the uniform and Gaussian model can be achieved by using the same equations as used in Chapter 5 [with the substitution of $\Sigma_n(\cdot) \rightarrow \Sigma_\omega\ N_\omega(\cdot)$], leading to Eqs. (6.3-5) and (6.3-6). Equations for the pole models are derived by a procedure similar to Chapter 5, Section 5.6. It was shown there that iterative maximization of Shannon–Einsteinian similarity

$$\max_{\mathbf{S}} \left\{ \sum_{\omega} N_\omega \ln F(\omega) \right\} \tag{A6-1}$$

over the entire set of model parameters $\{\mathbf{S}_k\}$ can be achieved by maximizing

$$\sum_{\omega} N_\omega\, f(k|\omega) \max_{\mathbf{S}_k} [\ln F(\omega|k)] \tag{A6-2}$$

at each iteration over the parameters of the kth source, $\mathbf{S}_k$, within the square brackets

[ln $F(\omega|k)$], while considering $f(k|\omega)$ known from the previous iteration. Thus it is sufficient to consider the following equations:

$$\sum_{\omega} N_{\omega}\, f(k|\omega)\ (\partial/\partial \mathbf{S}_k)\,[\ln\ F(\omega|k)] = 0 \tag{A6-3}$$

For the first-order pole model, the derivatives with respect to parameters ω_k and α_k are as follows:

$$(\partial/\partial\omega_k)\ \ln\ F(\omega|k) = 2\,(\omega - \omega_k)\left[(\omega - \omega_k)^2 + \alpha_k^2\right]^{-1} \tag{A6-4}$$

$$(\partial/\partial\alpha_k)\ \ln\ F(\omega|k) = \alpha_k^{-1} + 2\,\alpha_k\left[(\omega - \omega_k)^2 + \alpha_k^2\right]^{-1} \tag{A6-5}$$

The procedure in Section 5.6 calls for solving this joint system of equations for ω_k and α_k at every iteration. Instead, Eqs. (6.3-7) and (6.3-8) are derived by approximating this procedure: by considering the term $\left[(\omega - \omega_k)^2 + \alpha_k^2\right]$ to be known from the previous iteration. Equations for the second-order pole model parameters ω_k and α_k are derived by relating these parameters to the first two moments of the $F(\omega|k)$, leading to Eqs. (6.3-9) and (6.3-10). I did not prove that these modified procedures always converge; however, in my experience they did. If convergence problems are experienced, it is always possible to revert to the general Eq. (6.3-3).

NOTES

1. It is important not to confuse spatial, temporal, and frequency dimensions with the dimensions of a decision space, which was a subject of many examples in Chapter 5. For example, if an object occupies N color pixels, the decision space characterizing this object may include $4N$ dimensions: four measurements for each pixel (two spatial positions, one frequency, and one brightness). Psychological perception of color uses a nonlinear map from a continuous frequency into three basic colors. We will not be concerned in this book with psychological color perception.
2. For simplicity of notations, we use same letters to denote signals and their spectral components, while indicating the time or frequency domain by the coordinate t or ω; e.g., $s(t)$ and $S(\omega)$. This is especially convenient later, when we consider two-dimensional representations $S(t, \omega)$. Also, we use the same notation $S(\omega)$ for the true signal frequency components and for various ways in which they could be estimated. Usually, the context is sufficient to eliminate any ambiguity, otherwise, an explanation is added.
3. A similar effect is known in quantum electrodynamics (QED) under the name of the infrared catastrophe. It is known from QED that there is no "true" infinity of the number of emitted photons at $\omega \to 0$. This is because the lowest photon frequency is not zero, but is limited by the spatiotemporal extent of the emitting and measurement systems. Exact treatment of this issue in QED leads to logarithmic corrections $\sim \ln(\omega_{\min})$. This logarithmic dependence is weak, and our approximate treatment of this problem should be sufficient.

BIBLIOGRAPHICAL NOTES

Sunspot number modeling (Yule, 1927).
Brownian motion (Einstein; Wiener, 1948).

Stochastic stationary signals and classical spectrum estimation methods (Kay and Marple, 1981; Priestley, 1981; Proakis et al., 1992).

The first model-based neural network for signal processing was Widrow's Adaline (1959).

Parametric spectrum estimation methods (Kay and Marple, 1981; Proakis et al., 1992).

Einsteinian interpretation of the electromagnetic spectrum as proportional to a pdf of the photon frequency (Einstein and Hopf, 1910).

Efficient modeling and estimation procedures for spectrograms in this chapter (Perlovsky, 1994b, 1996b; Perlovsky et al., 1995b; 1997a,b).

Efficiency of adaptive algorithms (Widrow, 1988).

Classical AR estimators: the maximum likelihood (Yule, 1927; Walker, 1964; Levinson, 1947; Durbin, 1960)) and the Burg's maximum entropy (Burg, 1967).

Section 6.4 follows the development in Perlovsky (1994b).

Speech recognition reviews and further references (Zue, 1996; Olive et al., 1993).

HF clutter physical models: meteor and auroral causes of HF clutter (Vandrak et al., 1977) could not account for the properties of clutter observed around the globe. Phase Screen Ground Modulation (PSGM) model (Franchi and Tichovolsky, 1989). PSGM is based on the phase-screen model of Booker et al., (1987).

Previous approaches to clutter characterization were of a rather approximate nature, because of the difficulties related to multiple sources of radar signals interfering with estimation of properties of spread-Doppler clutter (Thomas, 1995).

PROBLEMS

6.1–1 Verify all steps leading from Eq. (6.1-17) to Eq. (6.1-17).

Step 1: By using a simple example (say $k = 1, 2; \omega = 1, 2, 3$), verify that the order of summation can be exchanged as follows [for any values of a_k, $s(\omega)$, $f(\omega, k)$]:

$$\sum_k a_k \sum_\omega s(\omega)\, f(\omega, k) = \sum_\omega s(\omega) \sum_k a_k\, f(\omega, k)$$

Hint: list all items in the double sums on each side. Thus, Eq. (6.1-16) can be rewritten as

$$\sum_\omega s(\omega) \sum_k a_k \exp(-i\omega t_{n-k}) = \sum_\omega \mathrm{wn}(\omega) \sum_l b_l \exp(-i\omega t_{n-l})$$

Step 2: By using (6.1-10), rewrite the above as

$$\sum_\omega s(\omega) \exp(-i\omega t_n) \sum_k a_k \exp(i\omega k \cdot \Delta)$$
$$= \sum_\omega \mathrm{wn}(\omega) \exp(-i\omega t_n) \sum_l b_l \exp(i\omega l \cdot \Delta)$$

Step 3: Both sides of the above equation are signals of t_n, which are defined as FTs. An FT is an orthogonal transformation: if signals are equal $[x(t) = y(t)]$, their FTs are also equal $[x(\omega) = y(\omega)]$. Compare each side of the above equation with the FT definition [Eq. (6.1-10)] and identify FTs on each side. These FTs are equal, yielding Eq. (6.1-17)

6.3–1 Verify normalization of Eq. (6.3-13). *Hint:* start with two-dimensional Gaussian pdf,

$$G(\mathbf{x}|k) = (2\pi)^{-1} (\det \mathbf{C}_k)^{-1/2} \exp\left\{-0.5 (\mathbf{x} - \mathbf{x}_k)^{\mathrm{T}} \mathbf{C}_k^{-1} (\mathbf{x} - \mathbf{x}_k)\right\}$$

and use Eqs. (6.2-7) and (6.2-11).

6.3–2 Define two-dimensional generalizations of ARMA models and derive estimation equations. First, consider a simple case of ARMA-frequency and uniform-time model. Then, consider ARMA-frequency and Gaussian-time model.

PREDICTION, TRACKING, AND DYNAMIC MODELS

In the beginning of every mythology, theology, or cosmogony there is a concept of the original chaos. An emergence of ordered cosmos is equated with the divine act of creation, which psychologically is equivalent to an emergence of consciousness. Thus, an ability to order and to predict is considered fundamental to consciousness. Originally, mathematical concepts of order were deterministic arithmetical and geometric ones. They have been seen as counteracting the mystery of chaos. Beginning with the sixteenth century, more sophisticated mathematical methods emerge, rationalizing the chaos itself. Prediction of outcomes in card games and gambling stimulated the development of probability theory. Early model-based approaches to prediction using linear regression and autoregressive modeling were used for more sophisticated, time-series prediction. Prediction methods got a significant boost on the one hand from the development of stochastic process theory, and on the other, from the need to solve the problem of target tracking that was first recognized during World War II, when radar was used to track aircraft. Today target tracking applications are numerous in both military and civilian areas of surveillance, navigation, guidance, air traffic control, and robotics. Sophisticated tracking methods are utilized for prediction, in particular, of stock markets. Stock market prediction replaced card games as a favorite breeding ground for new mathematical prediction methods: it serves to optimize worldwide investment and provides personal riches for those who can identify market inefficiencies. This section completes the overview of basic definitions, classical approaches, and relationships among prediction, tracking, and pattern recognition that began in Chapter 1, Section 1.3. New techniques are described for prediction and tracking in complicated situations of multiple concurrent processes, nonlinear relationships between variables, and when observations, in addition to signals of interest, contain noise and clutter (distracting signals). These new techniques are based on the modeling field theory and are extensions of MLANS and ENN model-based neural networks considered in the previous chapters.

7.1 Prediction, Association, and Nonlinear Regression

7.1.1 Multidimensional Linear Regression

In this section we formulate the classical linear regression technique for a case of multidimensional variables. We follow Sections 1.3.1 and 1.3.2, which considered a case of one-dimensional variables. Regression treats prediction as an estimation of unknown values of variables **y** from known values of variables **x**. Linear regression estimates a linear relationship between **x** and **y** from available past observations of pairs (**x**,**y**). A linear relationship between **x** and **y** can be estimated using two different approaches. The first approach starts with a model

$$\mathbf{y}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b} \tag{7.1-1}$$

where the matrix **A** and vector **b** are the parameters of the regression model that should be estimated from available data

$$\{(\mathbf{x}_n, \mathbf{y}_n), \quad n = 1, \ldots, N\} \tag{7.1-2}$$

This problem is solved as the ML estimation, by considering the likelihood function of the deviations of the model from the data

$$\varepsilon_n = \mathbf{y}_n - (\mathbf{A}\mathbf{x}_n + \mathbf{b}) \tag{7.1-3}$$

and by modeling the probability density of the deviations ε_n as a Gaussian function. The mathematical details of this approach are considered in Problem 7.1–1.

The second approach to estimating linear regression starts with a Gaussian model of the joint $\text{pdf}(\mathbf{x}, \mathbf{y}) = G(\mathbf{x}, \mathbf{y}|\mathbf{M}, \mathbf{C})$ and defines the regression of **y** on **x** as a conditional expected value of **y** given **x**,

$$\mathbf{y}(\mathbf{x}) = E\{\mathbf{y}|\mathbf{x}\} \equiv \int \mathbf{y}\ \text{pdf}(\mathbf{y}|\mathbf{x})\, d\mathbf{y}, \quad \text{pdf}(\mathbf{y}|\mathbf{x}) = G(\mathbf{x}, \mathbf{y})/\text{pdf}(\mathbf{x}) \tag{7.1-4}$$

Here, $\text{pdf}(\mathbf{y}|\mathbf{x})$ is a conditional density of **y**, given **x**. Mathematical details of this approach are given in Problem 7.1–2. The two approaches to the linear regression are equivalent for the following reason. A conditional probability above has a specific simple form for Gaussian densities: $\text{pdf}(\mathbf{x})$ and $\text{pdf}(\mathbf{y}|\mathbf{x})$ are Gaussian densities,

$$\text{pdf}(\mathbf{y}|\mathbf{x}) = G(\mathbf{x}, \mathbf{y}|\mathbf{M}, \mathbf{C})/G(\mathbf{x}|\mathbf{M}_x, \mathbf{C}_{xx}) = G\left[\mathbf{y}|\mathbf{M}'_y(\mathbf{x}), \mathbf{C}'_{yy}\right] \tag{7.1-5}$$

Substituting this into Eq. (7.1-4), we obtain, $\mathbf{y}(\mathbf{x}) = \mathbf{M}'_y(\mathbf{x})$.

Thus, both approaches to the linear regression yield the same result. Parameters of the regression Eq. (1.7-1) are given by

$$\mathbf{A} = \mathbf{C}_{yx}\ (\mathbf{C}_{xx})^{-1}, \quad \mathbf{b} = \overline{\mathbf{y}} - \mathbf{C}_{yx}\ (\mathbf{C}_{xx})^{-1}\, \overline{\mathbf{x}} \tag{7.1-6}$$

Here, mean values $\overline{\mathbf{x}}$, $\overline{\mathbf{y}}$ and covariances $\mathbf{C}_{xx}$, $\mathbf{C}_{yx}$ are estimated from the data:

$$\overline{\mathbf{x}} = (1/N) \sum_n \mathbf{x}_n, \quad \overline{\mathbf{y}} = (1/N) \sum_n \mathbf{y}_n \tag{7.1-7}$$

$$\mathbf{C}_{xx} = (1/N) \sum_n (\mathbf{x}_n - \overline{\mathbf{x}})^2, \quad \mathbf{C}_{yx} = (1/N) \sum_n (\mathbf{y}_n - \overline{\mathbf{y}}) (\mathbf{x}_n - \overline{\mathbf{x}})^{\mathrm{T}} \tag{7.1-8}$$

These parameters are related in a simple way to the parameters of joint pdf$(\mathbf{x}, \mathbf{y})$, the mean, $\mathbf{M}$, and covariance, $\mathbf{C}$,

$$\mathbf{M} = (\overline{\mathbf{x}}, \overline{\mathbf{y}}), \quad \mathbf{C} = \begin{Bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{Bmatrix} \tag{7.1-9}$$

where $\mathbf{C}_{xy} = \mathbf{C}_{yx}^{\mathrm{T}}$. It is more accurate to refer to $\mathbf{M}$ and $\mathbf{C}$ above as *estimates* of the parameters of pdf$(\mathbf{x}, \mathbf{y})$, obtained from data $\{(\mathbf{x}_n, \mathbf{y}_n)\}$.

Combining Eqs. (7.1-1) and (7.1-6), the regression is written by

$$\mathbf{y}(\mathbf{x}) = \overline{\mathbf{y}} + \mathbf{C}_{yx} (\mathbf{C}_{xx})^{-1} (\mathbf{x} - \overline{\mathbf{x}}) \tag{7.1-10}$$

This expression can be interpreted as follows: $\mathbf{y}(\mathbf{x})$ equals $\mathbf{y}$-average plus rescaled deviation of $\mathbf{x}$ from $\mathbf{x}$-average; the rescaling factor, $\mathbf{C}_{yx} (\mathbf{C}_{xx})^{-1}$, accounts for the correlation between $\mathbf{x}$ and $\mathbf{y}$ and for the difference in scales between $\mathbf{x}$ and $\mathbf{y}$.

7.1.2 Multidimensional Autoregression

Autoregression is an application of the linear regression method to time-series prediction. One-dimensional autoregression was introduced in Chapter 1, Section 1.3.3. Here we describe multidimensional autoregression. Also, in Section 1.3.3 the autoregressive model "memory" was limited to just one time-step back: x_t was predicted from x_{t-1}. Here, we derive equations for predicting $\mathbf{x}_t$ from $\mathbf{x}_{t-p}$, for $p = 1, \ldots, P$. At each time point t, the observation $\mathbf{x}_t$ is a D-dimensional vector, $\mathbf{x}_t = \left\{x_t^i, i = 1, \ldots, D\right\}$. And we consider a time series $\{\mathbf{x}_t, t = 1, \ldots, N\}$; for simplicity we use integer time values, but we use index t instead of our usual n, to emphasize the nature of the data as a time series. We also consider the data having zero-mean value; this can always be assured by subtracting the estimated mean, or by considering the differences of the original data, $\mathbf{x}_t \to (\mathbf{x}_t - \mathbf{x}_{t-1})$. The multidimensional autoregressive prediction model is

$$\mathbf{x}_t (\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-P}) = \sum_p \mathbf{A}_p \mathbf{x}_{t-p} \tag{7.1-11}$$

Parameters of this model are P matrixes, $\mathbf{A}_p$, each of $D \times D$ dimensions, $\mathbf{A}_p = \left\{A_p^{ij}\right\}$. These parameters can be estimated following the linear regression estimation procedure in the previous section. In this section, we should be a little careful with indexes i, j, and p (see Problem 7.1–3). The result of the estimation can be conveniently formulated by considering $\left\{A_p^{ij}\right\}$ as a set of vectors $\mathbf{a}^i$,

$$\mathbf{a}^i = \left\{a_l^i\right\}, \; a_l^i = A_p^{ij}, \quad l = j + (p-1) \cdot D \tag{7.1-12}$$

and by using a set of estimated correlation matrixes with different time lags p:

$$C_p^{ij} = (1/N) \sum_t x_t^i \, x_{t-p}^j, \quad C_{pp'}^{ij} = (1/N) \sum_t x_{t-p'}^i \, x_{t-p}^j \tag{7.1-13}$$

Using index l, we denote these matrixes as a set of vectors $\mathbf{c}^i$ and a matrix $\mathbf{C}$:

$$\mathbf{c}^i = \left\{c_l^i\right\} = \left\{C_p^{ij}\right\}, \quad \mathbf{C} = \{C_{ll'}\} = \left\{C_{pp'}^{jj'}\right\}, \quad l = j + (p-1)\cdot D,$$
$$l' = j' + (p'-1)\cdot D \tag{7.1-14}$$

With these notations, the parameters of the autoregressive model are given by

$$\mathbf{a}^i = \mathbf{C}^{-1}\,\mathbf{c}^i = \sum_{l'} C_{ll'}^{-1}\, c_l^i \tag{7.1-15}$$

Autoregression is a useful technique, when statistics of the process, that is matrixes **C**, do not change with time, or change slowly (stationarity assumption). Also, the classical autoregressive model, as well as the classical regression model in general, assumes that there is a single deterministic process determining the future mean value of the data (say, future prices), given by Eq. (7.1-1) or (7.1-11), and that other effects are random deviations from the mean. Assumption of the Gaussian density of the deviations further restricts adaptivity to linear combinations of inputs. But the stock market is not linear, and it is affected by a number of dynamic processes or forces acting concurrently. These limitations are overcome in the next section.

7.1.3 Nonlinear General Fuzzy Regression ANS (GFRANS)

Consider data $\{\mathbf{x}_n, \mathbf{y}_n, n = 1, \ldots, N\}$ as coming from multiple sources, $m = 1, \ldots, M$. For each source, assume a linear model relating $\mathbf{x}$ and $\mathbf{y}$,

$$\mathbf{y}_m(\mathbf{x}) = \mathbf{A}_m\mathbf{x} + \mathbf{b}_m \tag{7.1-16}$$

and the deviations from this model being random and Gaussian,

$$\text{pdf}(\mathbf{y}|\mathbf{x}, m) = G\left[\mathbf{y}|\mathbf{M}_y = \mathbf{y}_m(\mathbf{x}), \mathbf{C}_{ym}\right] \tag{7.1-17}$$

The likelihood of a data point $(\mathbf{x}, \mathbf{y})$ conditional on source m (that is, if we knew it came from source m), is the joint $\text{pdf}(\mathbf{x}, \mathbf{y}|m) = \text{pdf}(\mathbf{x}|m)\text{pdf}(\mathbf{y}|\mathbf{x}, m)$. Assuming a Gaussian density for $\text{pdf}(\mathbf{x}|m)$,

$$\text{pdf}(\mathbf{x}, \mathbf{y}|m) = G\left(\mathbf{x}|\mathbf{M}_x, \mathbf{C}_{xm}\right)\, G\left[\mathbf{y}|\mathbf{M}_y = \mathbf{y}_m(\mathbf{x}), \mathbf{C}_{ym}\right] \tag{7.1-18}$$

According to Section 7.1.1, this pdf is a Gaussian density, $\text{pdf}(\mathbf{x}, \mathbf{y}|m) = G(\mathbf{x}, \mathbf{y}|m)$, and every Gaussian density can be written in this form. Therefore, assumptions (7.1-16), (7.1-17), and (7.1-18) are equivalent to assuming a Gaussian joint $\text{pdf}(\mathbf{x}, \mathbf{y}|m)$ for each source.

In reality, we do not know which source is responsible for which observation, therefore, the probabilistic model for a data point is a combination of the alternatives,

$$\text{pdf}(\mathbf{x}, \mathbf{y}) = \sum_m r_m\, G(\mathbf{x}, \mathbf{y}|m) \tag{7.1-19}$$

The total log likelihood of the observed data, $\{\mathbf{x}_n, \mathbf{y}_n, n = 1, \ldots, N\}$ is given by

$$\text{LL} = \sum_n \ln\, \text{pdf}\,(\mathbf{x}_n, \mathbf{y}_n) = \sum_n \ln\left\{\sum_m r_m\, G\,(\mathbf{x}_n, \mathbf{y}_n|m)\right\} \tag{7.1-20}$$

This is a particular case of the general expression for AZ-similarity, Eq. (4.1-15), in which partial similarities $l(n|m)$ corresponding to alternative sources of data are given by Gaussian densities, $l(n|m) = r_m \text{ pdf } (\mathbf{x}_n, \mathbf{y}_n|m)$. The estimation of the parameters of these models is equivalent to estimating parameters of the mixture model $\{r_m, \mathbf{M}_m, \mathbf{C}_m\}$, which is performed by MLANS and was considered in detail in Chapter 5.

Let us derive the regression of $\mathbf{y}$ on $\mathbf{x}$, that is, the expected value of $\mathbf{y}$ given $\mathbf{x}$, for the above model. By combining Eqs. (7.1-4), (7.1-15), and (7.1-18),

$$\mathbf{y}(\mathbf{x}) = E\{\mathbf{y}|\mathbf{x}\} \equiv \int \mathbf{y} \text{ pdf}(\mathbf{y}|\mathbf{x})\, d\mathbf{y} = \sum_m r_m \int \mathbf{y} \,[\text{pdf}(\mathbf{x}, \mathbf{y}|m)/\text{pdf}(\mathbf{x})]\, d\mathbf{y} \tag{7.1-21}$$

Substitute, $\text{pdf}(\mathbf{x}, \mathbf{y}|m) = \text{pdf}(\mathbf{x}|m) \text{ pdf}(\mathbf{y}|\mathbf{x}, m)$:

$$\mathbf{y}(\mathbf{x}) = \sum_m r_m \,\text{pdf}(\mathbf{x}|m)/\text{pdf}(\mathbf{x}) \int \mathbf{y} \text{ pdf}(\mathbf{y}|\mathbf{x}, m)\, d\mathbf{y} \tag{7.1-22}$$

The integral here is just a definition of the mean value of $\mathbf{y}$ for $\text{pdf}(\mathbf{y}|\mathbf{x}, m)$, which is, according to Eq. (7.1-17), $\mathbf{y}_m(\mathbf{x})$. The term in front of the integral, $P(m|\mathbf{x}) = r_m \text{pdf}(\mathbf{x}|m)/\text{pdf}(\mathbf{x})$ is the a posteriori Bayes probability of the process m, given data $\mathbf{x}$. We obtain

$$\mathbf{y}(\mathbf{x}) = \sum_m P(m|\mathbf{x})\, \mathbf{y}_m(\mathbf{x}) \tag{7.1-23}$$

The a posteriori probabilities, $P(m|\mathbf{x})$, are simply expressed in terms of the parameters of the joint density $\text{pdf}(\mathbf{x}, \mathbf{y}|m)$. This is considered in Problem (7.1–6); the result is given simply by using the $\mathbf{x}$-component means and covariances of the joint density, $(\mathbf{M}_{mx}, \mathbf{C}_{mxx})$,

$$\begin{gathered} \text{pdf}(\mathbf{x}|m) = G\,(\mathbf{x}|\mathbf{M}_{mx}, \mathbf{C}_{mxx})\,, \quad \text{pdf}(\mathbf{x}) = \sum_m r_m G\,(\mathbf{x}|\mathbf{M}_{mx}, \mathbf{C}_{mxx}) \quad \text{and} \\ P(m|\mathbf{x}) = r_m G\,(\mathbf{x}|\mathbf{M}_{mx}, \mathbf{C}_{mxx}) \,/ \sum_{m'} r_{m'} G\,(\mathbf{x}|\mathbf{M}_{m'x}, \mathbf{C}_{m'xx}) \end{gathered} \tag{7.1-24}$$

So the nonlinear regression is obtained as a weighted sum of linear conditional regressions corresponding to various processes. And the weights are the a posteriori probabilities of the corresponding processes. In a part of $\mathbf{x}$-space, where one of the processes, m', is isolated from others [their pdfs do not overlap, $\text{pdf}(\mathbf{x}) \approx r_{m'} G\,(\mathbf{x}|\mathbf{M}_{m'x}, \mathbf{C}_{m'xx})$], $P(m'|\mathbf{x}) = 1$, and in this region, the regression is a linear function, the mth source conditional regression $\mathbf{y}(\mathbf{x}) = \mathbf{y}_m(\mathbf{x})$. But in general, the probabilities depend on $\mathbf{x}$, and the overall expression is a nonlinear function of $\mathbf{x}$.

For a practical utilization of the above nonlinear regression equation, it is necessary to estimate parameters of the joint density $\text{pdf}(\mathbf{x}, \mathbf{y})$ from available data $\{(\mathbf{x}_n, \mathbf{y}_n)\,,\ n = 1, \ldots, N\}$, using the Gaussian mixture model. This estimation problem was considered in details in Chapter 5, Eqs. (5.2-7) through (5.2-10). As was discussed several times before, the Gaussian mixture can model any probability density. Therefore, the nonlinear regression model derived in this section is the general model, suitable for modeling any relationships among data. A neural network implementation of this model is similar to MLANS considered in Chapter 5. Because of the fuzzy combination of multiple linear

regression models in Eq. (7.1-23), we call this neural network General Fuzzy Regression ANS, or GFRANS.

7.1.4 Nonlinear Autoregression

A regression model, when applied to modeling time-series data, results in an autoregressive model. Here, we derive the general nonlinear autoregressive model by combining results of two previous Sections 7.1.2 and 7.1.3. Again, this model can be interpreted as considering time-series data, $\{\mathbf{x}_t, t = 1, \ldots, N\}$, governed by several linear autoregressive processes, $m = 1, \ldots, M$. We would like to predict $\mathbf{Y} = \mathbf{x}_t$ from $\mathbf{X} = (\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-P})$, therefore, similar to the regression model, we consider a joint density of $(\mathbf{Y}, \mathbf{X}) = (\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-P})$, and we model it using a Gaussian mixture density. Each component of the mixture describes the mth data source,

$$G(\mathbf{X}, \mathbf{Y}|m) = G\left(\mathbf{X}, \mathbf{Y}|\mathbf{M}_m, \mathbf{C}_m\right), \quad \mathbf{M}_m = \left(\mathbf{M}_{mx}, \mathbf{M}_{my}\right),$$
$$\mathbf{M}_{my} = \sum_p \mathbf{A}_{mp}\mathbf{x}_{t-p} \tag{7.1-25}$$

The mean $\mathbf{M}_{my}$ is time dependent, so we will denote it $\mathbf{M}_{myt}$. Covariances $\mathbf{C}_m$ are considered to be constant parameters. Thus, the parameters of the model, $\mathbf{S}_m$, are

$$\mathbf{S}_m = \left\{r_m, \mathbf{M}_{mx}, \mathbf{A}_{mp}, \mathbf{C}_m\right\} \tag{7.1-26}$$

These parameters should be estimated from the available past data $\{\mathbf{x}_t, t = 1, \ldots, N\}$. The estimation equations are derived from the general MFT equations of Chapter 4 or by combining the derivations in Chapter 5 and Section 7.1.2. This derivation is considered in Problem 7.1–8, and here we summarize the results. [For simplicity, we consider the case of $\mathbf{M}_{mx} = 0$; this is usually applicable for the difference time series, $(\mathbf{x}_t - \mathbf{x}_{t-1})$, even if not applicable for the original data $\mathbf{x}_t$. Modifications of this result for the case of $\mathbf{M}_{mx} \neq 0$ are considered in Problem 7.1–9.] The results are intuitively appealing; they consist of a simple modification of the equations in Section 7.1-7: sums over data are changed into weighted sums with weights being the a posteriori probabilities for each source. Equation (7.1-13) is changed into

$$C^{ij}_{p,m} = (1/N_m) \sum_t P(m|t) x^i_t x^j_{t-p}, \quad C^{ij}_{pp',m} = (1/N_m) \sum_t P(m|t) x^i_{t-p'} x^j_{t-p} \tag{7.1-27}$$

where

$$P(m|t) = r_m G\left(\mathbf{X}, \mathbf{Y}|\mathbf{M}_m, \mathbf{C}_m\right) / \sum_{m'} r_{m'} G\left(\mathbf{X}, \mathbf{Y}|\mathbf{M}_{m'}, \mathbf{C}_{m'}\right) \tag{7.1-28}$$

and

$$N_m = \sum_t P(m|t) \tag{7.1-29}$$

Equation (7.1-27) describes $\mathbf{C}_{mxx}$ and $\mathbf{C}_{mxy}$ components of matrixes $\mathbf{C}_m$. The complete matrixes $\mathbf{C}_m$ dimensions $(P + 1) \times (P + 1)$ are given by the second part of Eq. (7.1-27)

considered for $p, p' = 0, \ldots, P$. Similar to Eq. (7.1-14), we use index l to denote **C**-matrixes as a set of vectors $\mathbf{c}_m^i$ and matrixes $\mathbf{C}_{mxx}$:

$$\mathbf{c}_m^i = \left\{c_{l,m}^i\right\} = \left\{C_{p,m}^{ij}\right\}, \quad \mathbf{C}_{mxx} = \left\{C_{ll',m}\right\} = \left\{C_{pp',m}^{jj'}\right\}$$
$$i, j, j' = 1, \ldots, D; \quad p, p' = 1, \ldots, P; \quad l = j + (p-1)\cdot D,\ l' = j' + (p'-1)\cdot D \tag{7.1-30}$$

With these notations, the parameters of the nonlinear autoregressive model are estimated using the following equations:

$$\mathbf{a}_m^i = \mathbf{C}_{mxx}^{-1}\mathbf{c}_m^i = \sum_{l'} C_{ll',m}^{-1}\, c_{l,m}^i, \quad r_m = N_m/N$$
$$\mathbf{C}_m = \left\{C_{pp',m}^{jj'}\right\}, \quad p, p' = 0, \ldots, P \tag{7.1-31}$$

Similar to the general MFT considered in Chapter 4 and to MLANS considered in Chapter 5, these equations are iterative estimation equations, since the a posteriori probabilities in the right-hand sides of these equations are functions of the parameters. In each iteration, parameter estimation is improved, followed by improved probability estimation.

The general nonlinear autoregressive equation is derived following the previous section (see Problem 7.1–10). It is a weighted sum of the linear autoregressive models for each source, with weights being the a posteriori probabilities of sources, given the available data $\mathbf{X}_{t-1} = (\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-P})$,

$$\mathbf{x}_t\,(\mathbf{X}_{t-1}) = \sum_m P\,(m|\mathbf{X}_{t-1})\,\mathbf{M}_{myt} = \sum_m P\,(m|\mathbf{X}_{t-1}) \sum_p \mathbf{A}_{mp}\mathbf{x}_{t-p}$$
$$P\,(m|\mathbf{X}_{t-1}) = r_m G\,(\mathbf{X}_{t-1}|\mathbf{M}_{mx}, \mathbf{C}_{mxx}) \,/ \sum_{m'} r_{m'} G\,(\mathbf{X}_{t-1}|\mathbf{M}_{m'x}, \mathbf{C}_{m'xx}) \tag{7.1-32}$$

Similar to the general nonlinear regression, the nonlinear autoregression is obtained as a weighted sum of linear conditional autoregressions corresponding to various processes. And the weights are the a posteriori probabilities of the corresponding processes. In a part of $(\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-P})$-space, where one of the processes, m', is isolated from others (their pdfs do not overlap), $P\left(m'|\mathbf{X}_{t-1}\right) = 1$, and in this region, the autoregression is a linear function, the mth source conditional autoregression $\mathbf{M}_{myt}$. But in general, the probabilities depend on $\mathbf{X}_{t-1}$, and the overall expression is a nonlinear function of $\mathbf{X}_{t-1}$. Since the Gaussian mixture can model any probability density, the nonlinear autoregressive model derived in this section is the general model, suitable for modeling any relationships among time-series data.

7.1.5 Example: Data Mining and Revenue Prediction

Let us illustrate the developed theory with an application example. In this application, it was necessary to identify various factors and drivers that determine hospital revenues. To accomplish this it was necessary to develop a predictive model relating revenues at the patient level to the demographic, clinical, and other factors that are present in publicly available databases. The difficulty of the problem was due to the existence of multiple

"forces" determining the relationships between the revenue and other data. Because of this, the relationships are nonlinear: the same factors are related in a different way for different patients. The complicated nonlinear relationship makes classical linear regression inapplicable to this problem. The GFRANS technique described in Section 7.1.3 was utilized. GFRANS establishes groupings among patients, corresponding to the "forces," while determining the relationships among various factors and the revenue. The predicted variable y was the revenue, and the data characterizing the patient, $\mathbf{x}$, was selected from the databases. This selection was a result of preliminary analysis using GFRANS, with an in-depth examination of the different groups into which the data were segmented and comparison of various revenue prediction performance results. This analysis indicated that eight factors or features were the dominant ones for estimating the revenues; these features were combined into the eight-dimensional data vector $\mathbf{x}$ used for the final analysis.

The analyzed databases contained more than 100,000 cases from more than 50 hospitals in the state of Florida. GFRANS identified 16 groups of cases corresponding to 16 "forces" determining the relationships between the revenues and other factors. These groups indicated geographic differences as well as differences due to varying arrangements between hospitals and payors (payors are insurance companies, HMOs, and Government organizations that make payments on behalf of individual patients). The results are illustrated in Figure 7.1-1 by showing the histograms of deviations between the predictive model and the actual data. Figure 7.1-1a shows results of linear regression and Fig. 7.1-1b shows results of GFRANS. A dramatic improvement in accuracy is obvious.

7.1.6 Summary of Section 7.1

Powerful nonlinear prediction models have been derived in this section, which are closely related to the MLANS estimation technique. The general nonlinear regression and autoregressive models are applicable to characterizing any types of data, due to the fact that

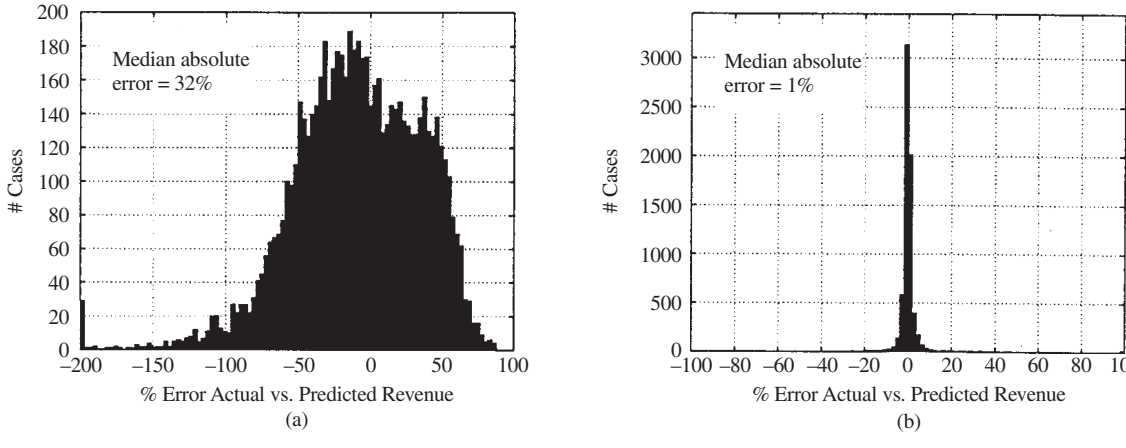


Figure 7.1-1 Performance of linear regression model vs. GFRANS is illustrated by using error histograms, which show the deviations between the model and actual revenues; **(a)** linear regression model, **(b)** GFRANS.

Gaussian mixtures can model any pdf. Their practical utility is greatest when data are characterized by several sources or processes, and each source is Gaussian, that is, each source can be characterized by a linear model. While estimating models, the developed technique performs two functions concurrently: it associates data with each source and estimates parameters of the sources. The association is fuzzy; each data sample is partitioned in a fuzzy way among all sources. At the beginning of MLANS iterations, the associations are highly fuzzy; in the process of learning, the fuzziness is reduced and converges to probabilistic or crisp association. Thus, MLANS "recognition" of signal sources resembles Aristotelian Forms converging to definite concepts.

7.2 Association and Tracking Using Bayesian MFT

Object tracking concerns detection of moving objects and estimation of their trajectories in sensory data. Data originate from multiple sources: objects of interest called targets, objects of no interest called background or clutter, and sensor noise. Presence of multiple sources makes tracking similar to nonlinear regression considered above: sources of signals have to be separated, while parameters of each source are being estimated. The tracking problem has been traditionally approached by subdividing it into several simpler steps, also called surveillance functions: detection, association, and track estimation. Detection refers to the process of determining samples or pixels of data containing target signals while rejecting clutter and noise. Association refers to grouping of data from multiple frames into subsets corresponding to a single object. And track estimation refers to estimating parameters of the models of object motion (position, velocity, etc.). When surveillance functions can be performed sequentially, that is, detection first, association second, and track estimation third, classical approaches to tracking are efficient. These approaches were overviewed in Chapter 1, Section 1.3.

But when target signals are of the same order of magnitude as clutter signals, target detection cannot be performed on a single frame or scan. Multiple time measurements have to be utilized for target detection, which requires knowledge of tracks. Thus the problems of detection, association, and track estimation have to be solved concurrently. Such a capability is referred to as track-before-detect or, more accurately, as Concurrent Association and Tracking (CAT). Classical approaches to CAT require associating multiple subsets of data with multiple possible track models, which often leads to a combinatorial explosion of complexity. Mathematical formulation of this general tracking problem resembles that of model-based vision. And this is not surprising; CAT approaches tracking as recognition of spatiotemporal patterns. This section describes an approach to CAT based on MFT that solves the problem of combinatorial complexity. We describe both variants of MFT, Bayesian MLANS and Shannon's ENN, and present application examples.

7.2.1 Concurrent Association and Tracking (CAT)

Consider a sequence of visual images $I(\mathbf{x}, t)$, where I is image intensity, and $(\mathbf{x}, t)$ designates spatial and temporal coordinates: $\mathbf{x} = (x, y)$ are pixels, corresponding to the two angular coordinates and t are time frames. For color images, intensity is a vector of red,

green, and blue signals, $\mathbf{I}$. Often, we will number pixels by index $n = 1, \ldots, N$, $(\mathbf{x}_n, t_n) = (x_n, y_n, t_n)$, $\mathbf{I}(\mathbf{x}, t) = \mathbf{I}_n$; this will remind us that $(\mathbf{x}, t)$ are not just pixel indexes, but also physical coordinates of sources of signals, so that pixel n is characterized by a set of measurements $\mathbf{z}_n = (x_n, y_n, t_n, \mathbf{I}_n)$. Most often we will consider coordinates and intensities as functions of time, so it will be convenient to denote nth pixel measurements as

$$\mathbf{z}_n = \mathbf{z}(t_n) = (x_n, y_n, \mathbf{I}_n) \tag{7.2-1}$$

Each pixel measurement is produced by an object, or, possibly by multiple objects (in a case of unresolved objects), numbered $k = 1, \ldots, K$. First, we consider the Bayesian-MFT approach to tracking; Shannon–Einsteinian formulation is considered later. In Bayesian-MFT, the deterministic part of a model describes the predicted expected value of the measurements, conditional on a particular type of object, k:

$$\mathbf{M}_{nk}(\mathbf{S}_k, t_n) = E\{\mathbf{z}_n | H_k\} \tag{7.2-2}$$

Here, $\mathbf{S}_k$ are model parameters that in tracking applications are usually called state parameters, and H_k are hypotheses of the type of object, its motion model, and parameters, etc. For example, an unresolved object with a constant intensity, $\mathbf{I}_k$, moving with a constant velocity[1] $\mathbf{V}_k$, is modeled as

$$\mathbf{M}_{nk}(\mathbf{S}_k, t_n) = E\{(\mathbf{x}_n, \mathbf{I}_n) | H_k\} = (\mathbf{R}_k + \mathbf{V}_k t_n, \mathbf{I}_k) \tag{7.2-3}$$

State parameters $\mathbf{S}_k = (\mathbf{R}_k, \mathbf{V}_k, \mathbf{I}_k)$, where $\mathbf{R}_k$ is object position at time $t_n = 0$.

Random uncertainty about target position, such as due to sensor errors, is modeled probabilistically, by considering pdf $(\mathbf{z}_n | H_k)$. Let us first consider Gaussian densities to model conditional pdfs:

$$\text{pdf}(\mathbf{z}_n | k) \equiv \text{pdf}(\mathbf{z}_n | H_k) = G[\mathbf{z}_n | \mathbf{M}_{nk}(\mathbf{S}_k, t_n), \mathbf{C}_{nk}(\mathbf{S}_k, t_n)] \tag{7.2-4}$$

and, for simplicity, we consider covariance matrixes to be constant for every type of track $\mathbf{C}_{nk}(\mathbf{S}_k, t_n) = \mathbf{C}_k$. Since we do not know which object produces the pixel-n signal, the probabilistic model of the signal is a sum over alternative sources/objects

$$\text{pdf}(\mathbf{z}_n) = \sum_k r_k G(\mathbf{z}_n | k) = \sum_k r_k G(\mathbf{z}_n | \mathbf{M}_{nk}(\mathbf{S}_k, t_n), \mathbf{C}_k) \tag{7.2-5}$$

The total log likelihood of the observed data, $\{\mathbf{z}_n, n = 1, \ldots, N\}$ is given by

$$\text{LL} = \sum_n \ln \text{pdf}(\mathbf{z}_n) = \sum_n \ln \left\{ \sum_k r_k G(\mathbf{z}_n | k) \right\} \tag{7.2-6}$$

This is a particular case of the general expression for AZ-similarity, Eq. (4.1-15), in which partial similarities $l(n|k)$ corresponding to alternative sources of data (objects) are given by Gaussian densities, $l(n|k) = r_k G(\mathbf{z}_n | k)$. The estimation of the state parameters can be performed using the general MFT equations for complex models, or MLANS-type equations for more simple models. MFT and MLANS perform concurrent association of data with object tracks ($\mathbf{z}_n$ with k) and estimation of model parameters, $\mathbf{S}_k$.

Below, we consider several models useful for tracking and derive MLANS-type CAT estimation equations.

7.2.2 Linear Model for Tracking

Consider an unresolved object with a constant intensity, $\mathbf{I}_k$, moving with a constant velocity $\mathbf{V}_k$ (see Note 1). It is modeled as

$$\mathbf{M}_{nk}\left(\mathbf{S}_k, t_n\right) = \left(\mathbf{X}_{kn}, \mathbf{I}_k\right), \quad \mathbf{X}_{kn} = \mathbf{X}_k\left(t_n\right) = \mathbf{R}_k + \mathbf{V}_k\, t_n \tag{7.2-7}$$

where $\mathbf{R}_k$ is the object position at time $t_n = 0$. [If a reference point of t_0 is desirable, it is sufficient to substitute in the above equation $t_n \to (t_n - t_0)$.]. Consider random deviations of the measured intensity from its mean value to be independent from random deviations in the object position.[2] Then, the object's pdf is modeled as a product of intensity and dynamic-model pdfs. This results in the object-k conditional pdf given by

$$\text{pdf}\left(\mathbf{z}_n|H_k\right) = G\left(\mathbf{I}_n|k\right) G\left(\mathbf{x}_n|k\right) = G\left(\mathbf{I}_n|\mathbf{I}_k, \mathbf{CI}_k\right) G\left(\mathbf{x}_n|\mathbf{X}_{kn}, \mathbf{CX}_k\right) \tag{7.2-8}$$

In this section on tracking, we often use the same variable names with different indexes to denote data, $\mathbf{x}_n, \mathbf{I}_n$, and their models, $\mathbf{x}_k, \mathbf{I}_k$. State parameters of this model include r_k, the object-rate (that is, the proportion of object pixels among the total number of pixels), $\mathbf{S}_k = (r_k, \mathbf{R}_k, \mathbf{V}_k, \mathbf{I}_k, \mathbf{CI}_k, \mathbf{CX}_k)$. The CAT estimation equations for $r_k, \mathbf{I}_k, \mathbf{CI}_k, \mathbf{CX}_k$, are similar to those derived previously for Gaussian mixtures (see Problem 7.2–1). We write these equations here, using bracket notations for neuronal operations; for any quantity f_n, its weighted sum

$$\langle f_n \rangle = \sum_n W_{nk} f_n, \quad W_{nk} = P(k|n) = r_k \text{pdf}\left(\mathbf{z}_n|k\right) / \sum_{k'} r_{k'} \text{pdf}\left(\mathbf{z}_n|k'\right) \tag{7.2-9}$$

With these notations,

$$\begin{aligned} r_k &= N_k/N,\ N_k = \langle 1 \rangle \\ \mathbf{I}_k &= \langle \mathbf{I}_n \rangle / \langle 1 \rangle \\ \mathbf{CI}_k &= \left\langle \left(\mathbf{I}_n - \mathbf{I}_k\right)\left(\mathbf{I}_n - \mathbf{I}_k\right)^{\mathrm{T}} \right\rangle / \langle 1 \rangle \\ \mathbf{CX}_k &= \left\langle \left(\mathbf{x}_n - \mathbf{X}_{kn}\right)\left(\mathbf{x}_n - \mathbf{X}_{kn}\right)^{\mathrm{T}} \right\rangle / \langle 1 \rangle \end{aligned} \tag{7.2-10}$$

The CAT estimation equations for $\mathbf{R}_k, \mathbf{V}_k$, are derived in Problem 7.2–1. It is a system of two linear equations for the parameters of each object-track:

$$\begin{aligned} \langle \mathbf{x}_n \rangle - \mathbf{R}_k \langle 1 \rangle - \mathbf{V}_k \langle t_n \rangle &= 0 \\ \langle \mathbf{x}_n t_n \rangle - \mathbf{R}_k \langle t_n \rangle - \mathbf{V}_k \left\langle t_n^2 \right\rangle &= 0 \end{aligned} \tag{7.2-11}$$

From here,

$$\begin{aligned} \mathbf{R}_k &= \left[\langle \mathbf{x}_n \rangle \left\langle t_n^2 \right\rangle - \langle \mathbf{x}_n t_n \rangle \langle t_n \rangle\right] / \det \\ \mathbf{V}_k &= \left[\langle \mathbf{x}_n t_n \rangle \langle 1 \rangle - \langle \mathbf{x}_n \rangle \langle t_n \rangle\right] / \det \\ \det &= \left[\langle 1 \rangle \left\langle t_n^2 \right\rangle - \langle t_n \rangle \langle t_n \rangle\right] \end{aligned} \tag{7.2-12}$$

where det is a determinant of the system of Eqs. (7.2-11). Equations (7.2-9) through (7.2-12) define an iterative procedure for concurrent association and estimation of the model parameters.

7.2.3 Second-Order Model for Tracking

For turning or accelerating objects, the second-order track models, which account for acceleration, could be more appropriate. The state parameters for the second-order model include acceleration, $\mathbf{A}_k$, and the dynamic part of the model is given by

$$\mathbf{X}_{kn} = \mathbf{X}_k(t_n) = \mathbf{R}_k + \mathbf{V}_k t_n + 0.5\,\mathbf{A}_k\, t_n^2 \tag{7.2-13}$$

The MLANS CAT equations for the state parameters of the second-order track model are derived in a manner similar to the above (see Problem 7.2–2). Equations (7.2-10) for r_k, $\mathbf{I}_k$, $\mathbf{CI}_k$, $\mathbf{CX}_k$ do not change. Equations (7.2-11) become a system of three linear equations for the parameters of each class:

$$\begin{aligned} \langle \mathbf{x}_n \rangle - \mathbf{R}_k \langle 1 \rangle - \mathbf{V}_k \langle t_n \rangle - 0.5\,\mathbf{A}_k \left\langle t_n^2 \right\rangle &= 0 \\ \langle \mathbf{x}_n t_n \rangle - \mathbf{R}_k \langle t_n \rangle - \mathbf{V}_k \left\langle t_n^2 \right\rangle - 0.5\,\mathbf{A}_k \left\langle t_n^3 \right\rangle &= 0 \\ \left\langle \mathbf{x}_n t_n^2 \right\rangle - \mathbf{R}_k \left\langle t_n^2 \right\rangle - \mathbf{V}_k \left\langle t_n^3 \right\rangle - 0.5\,\mathbf{A}_k \left\langle t_n^4 \right\rangle &= 0 \end{aligned} \tag{7.2-14}$$

leading to the following CAT equations:

$$\mathbf{R}_k = \mathrm{det1}/\,\mathrm{det}, \quad \mathbf{V}_k = \mathrm{det2}/\,\mathrm{det}, \quad \mathbf{A}_k = \mathrm{det3}/\,\mathrm{det} \tag{7.2-15}$$

where det, det1, det2, and det3 are the corresponding determinants of the system of Eqs. (7.2-14). Instead of Eq. (7.2-15), one can use a standard subroutine to solve the system of Eqs. (7.2-14) at each iteration.

7.2.4 Link-Track Model

The link-track model is designed to model maneuvering objects. It is a piecewise linear track, composed of linear links. It differs from consecutive applications of the linear track model in two aspects: (1) links are constrained by the requirement of a connected track without jumps or gaps (that is, the last position of each link is the first position of the next link) and (2) as more data are acquired previous links can be reestimated if this results in a more likely overall track.

Individual links are numbered $l = 1, \ldots, L$, and link parameters of the object (track) k are numbered by two indexes (k, l). The link starting time is denoted as t_{kl}. The time extent of all links is identical and currently is specified by the operator. The model equation for each link is similar to the linear model Eq. (7.2-7)

$$\mathbf{X}_{kl}\,(t_n) = \mathbf{R}_{kl} + \mathbf{V}_{kl}\left(t_n - t_{k,l}\right) \tag{7.2-16}$$

The connectivity constraint is expressed mathematically as

$$\mathbf{X}_{k,l}\left(t_{k,l+1}\right) = \mathbf{X}_{k,l+1}\left(t_{k,l+1}\right), \quad l = 1, \ldots, L-1 \tag{7.2-17}$$

or, in terms of the model parameters,

$$\mathbf{R}_{k,l} + \mathbf{V}_{k,l} \cdot \left(t_{k,l+1} - t_{k,l}\right) = \mathbf{R}_{k,l+1}, \quad l = 1, \ldots, L-1 \tag{7.2-18}$$

When deriving MLANS CAT equations, these constraints are accounted for by the method of Lagrange multipliers, which introduces $(L-1)$ Lagrange coefficients for every track, $\lambda_{kl}, l = 2, \ldots, L$ (see Problem 7.2–3). The angular bracket notations for the link-track model below are modified by explicitly specifying the class and link for which they are computed, so we use $\langle t_n | kl \rangle$ instead of $\langle t_n \rangle$. With these notations, CAT equations for the parameters $\mathbf{R}_{k,l}$ and $\mathbf{V}_{k,l}$ are

$$\begin{gathered} \langle \mathbf{x}_n | kl \rangle - \mathbf{R}_{k,l} \langle 1 | kl \rangle - \mathbf{V}_{k,l} \left\langle t_n - t_{k,l} | kl \right\rangle + \lambda_{kl} - \lambda_{k,l+1} = 0 \\ \left\langle \mathbf{x}_n \left(t_n - t_{k,l}\right) | kl \right\rangle - \mathbf{R}_{k,l} \left\langle t_n - t_{k,l} | kl \right\rangle - \mathbf{V}_{k,l} \left\langle \left(t_n - t_{k,l}\right)^2 | kl \right\rangle - \left(t_n - t_{k,l}\right) \lambda_{k,l+1} = 0 \end{gathered} \tag{7.2-19}$$

Here, $l = 1, \ldots, L$, and $\lambda_{kl} = 0$ for $l = 1$ or $L+1$. Together, Eqs. (7.2-18) and (7.2-19) comprise a system of $3L-1$ linear equations for $3L-1$ unknowns $\left(\mathbf{R}_{k,l}, \mathbf{V}_{k,l}, \lambda_{k,l}\right)$. For a given MLANS iteration, the solution of these equations yields an estimate of parameters of the link-track model. Together, Eqs. (7.2-9), (7.2-10), (7.2-18), and (7.2-19) define an iterative procedure for estimation of model parameters. A standard subroutine can be used to solve the system of Eqs. (7.2-19) at each CAT MLANS iteration.[3]

7.2.5 Random Noise and Clutter Model

Localized and moving clutter objects are tracked using the same track models as targets that are described above. In addition, for distributed random clutter and noise signal sources that do not correspond to localized objects, it is often convenient to use a simple noise/clutter track model with a uniform density of the position vectors $\mathbf{x}_n$,

$$\text{pdf}\left(\mathbf{x}_n | H = \text{clutter}\right) = \left\{\left[x_{\max} - x_{\min}\right]\left[y_{\max} - y_{\min}\right]\right\}^{-1} = \Delta^{-1} \tag{7.2-20}$$

The estimation Eqs. (7.2-10) for parameters $r_k, \mathbf{I}_k, \mathbf{CI}_k, \mathbf{CX}_k$ do not change.

7.2.6 Active Sensor and Doppler Track Models

Track models considered above are suitable for sensors measuring positional coordinates of objects in their field of view. Passive sensors, such as a TV camera, measure two-dimensional angular positions of objects. Active sensors, such as radars, ladars, and sonars, usually measure range and one or two angular coordinates. Tracking in range and azimuth, or in three dimensions, requires very little modifications to the above models and estimation equations. In fact, the only modification required is to Eq. (7.2-20), when tracking in three dimensions. But in addition to positional coordinates, some active sensors also measure radial velocities, by using the Doppler effect. The Doppler effect is a change in frequency of reflected waves due to the relative radial motion of the sensor and reflecting object. Doppler measurements of velocity are usually much more accurate than estimation of velocity from positional measurements, using track models. When Doppler measurements are available, the above models should be modified. The modification is discussed here only

for the linear track model, and following this example, the reader can derive modifications to other models as needed.

Required modifications involve the model and estimation of radial object position and velocity, while angular motion models and estimation equations remain unchanged. We denote the range measurements r_n and Doppler velocity measurements v_n. The vector of these radial coordinate and velocity measurements (r_n, v_n) we denote by $\mathbf{xr}_n$, and its model by $\mathbf{XR}_{kn}$. The model for r_n is given by the radial component of Eq. (7.2-7),

$$X_{kn} = X_k(t_n) = R_k + V_k t_n \tag{7.2-21}$$

and the model for v_n is given by the radial component of the velocity model, V_k, so in the linear model, the range is a linear function of time, and velocity is a constant:

$$\mathbf{XR}_{kn} = \mathbf{XR}_k(t_n) = (R_k + V_k t_n, V_k) \tag{7.2-22}$$

The object-k conditional pdf is modeled as a Gaussian. For simplicity, we assume uncorrelated errors in r_n, v_n, and angular coordinates, then

$$\text{pdf}(\mathbf{xr}_n|H_k) = G(\mathbf{xr}_n|k) = G(r_n|(R_k + V_k t_n), CR_k)\ G(v_n|V_k, CV_k) \tag{7.2-23}$$

Following the previous derivations, the CAT estimation equations for R_k and V_k are derived in Problem 7.2–4. For all parameters except R_k and V_k, the equations are the same as in Section 7.2.2. For R_k and V_k we obtain a system of two linear equations for the parameters of each object-track:

$$\begin{aligned} \langle r_n \rangle - R_k \langle 1 \rangle - V_k \langle t_n \rangle &= 0 \\ (\langle r_n t_n \rangle + \alpha \langle v_n \rangle) - R_k \langle t_n \rangle - V_k \left(\left\langle t_n^2 \right\rangle + \alpha \langle 1 \rangle\right) &= 0 \end{aligned} \tag{7.2-24}$$

Here, $\alpha = CR_k / CV_k$. Terms proportional to α are large, so approximately, $V_k = \langle v_n \rangle / \langle 1 \rangle$.

7.2.7 Autoregressive Model for Tracking

In many cases, object motion can be modeled by autoregressive models. A second-order autoregressive model can be viewed as a motion with constant velocity between time t and $t + 1$, with random maneuvers (velocity changes) at each t. Similarly, a third-order autoregressive model can be viewed as a motion with constant acceleration between time t and $t + 1$, with random maneuvers (acceleration changes) at each t. There is a considerable degree of similarity between tracking and the nonlinear autoregression considered in Section 7.1.4, with each conditional autoregressive model corresponding to an object-track. Still, one cannot simply use the nonlinear autoregression estimation equations for tracking. The reason is that in Section 7.1.4, we considered the model at time t as a function of previous observations $\{\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-p}\}$, but in case of tracking, the autoregressive model should be applied to the previous estimated model values of the object position:

$$\mathbf{x}_{kn} = \mathbf{x}_k(t_n) = \sum_p \mathbf{A}_{kp} \mathbf{x}_k(t_n - p) \tag{7.2-25}$$

Since $\mathbf{x}_k(t_n - p)$ is a function of $\mathbf{A}_{kp'} \mathbf{x}_k\left(t_n - p - p'\right)$, and $\mathbf{x}_k\left(t_n - p - p'\right)$ is a function of $\mathbf{A}_{kp''} \mathbf{x}_k\left(t_n - p - p' - p''\right)$, etc., the derivative expands into a chain of expressions going

to $t = 0$. The derivation of neural equations is getting involved and is not considered here. Interested readers can derive these equations on their own; recursive neural equations for this type of problem are considered in Perlovsky and Jaskolski (1994).

7.2.8 Models for Tracking Resolved Objects

When tracking resolved objects, the only modifications that are needed involve accounting for the models of spatial distribution of object intensity. We consider here a simple case of tracking a circular object with constant intensity, using a linear track model. The geometric model is characterized by three parameters, object center $\mathbf{xc}_k\,(t_n)$, radius a_k, and intensity $\mathbf{I}_k$. The track model, $\mathbf{X}_{kn}$, describes the motion of the object center. It turns out that equations are much simplified if a slightly different geometric model is considered: instead of a circle, consider a geometrically fuzzy object, with the pixel-object membership function given by a Gaussian centered at $\mathbf{xc}_k\,(t_n)$ with standard deviation of a_k, $G\left[\mathbf{x}_n|\mathbf{xc}_k\,(t_n)\,, a_k^2\right]$. The conditional pdf, instead of Eq. (7.2-8), is given by

$$\text{pdf}\,(\mathbf{z}_n|H_k) = G\left[\mathbf{x}_n|\mathbf{xc}_k\,(t_n)\,, a_k^2\right]\,G\,(\mathbf{I}_n|\mathbf{I}_k, \mathbf{CI}_k)\;G\,[\mathbf{xc}_k\,(t_n)\,|\mathbf{X}_{kn}, \mathbf{CX}_k] \tag{7.2-26}$$

The estimation equations are derived in Problem 7.2–5; in place of Eq. (7.2-11), we obtain

$$\begin{aligned}\left\langle\left\langle \mathbf{x}_{n'}\delta_{tn,tn'}\right\rangle' \Big/ \left\langle\delta_{tn,tn'}\right\rangle'\right\rangle - \mathbf{R}_k\,\langle 1\rangle - \mathbf{V}_k\,\langle t_n\rangle = 0\\ \left\langle t_n\left\langle \mathbf{x}_{n'}\delta_{tn,tn'}\right\rangle' \Big/ \left\langle\delta_{tn,tn'}\right\rangle'\right\rangle - \mathbf{R}_k\,\langle t_n\rangle - \mathbf{V}_k\left\langle t_n^2\right\rangle = 0\end{aligned} \tag{7.2-27}$$

Here, $\langle f_{n'}\rangle' = \Sigma_{n'}\;P(k|n')\,(f_{n'})$, for any $f_{n'}$, so that $N_k\,(t_n) = \left\langle\delta_{tn,tn'}\right\rangle'$ is an estimated number of pixels in frame t_n, associated with object k, $\mathbf{x}_k\,(t_n) = \left\langle\mathbf{x}_{n'}\delta_{tn,tn'}\right\rangle' \Big/ \left\langle\delta_{tn,tn'}\right\rangle'$ is an estimated average position of these pixels, $\left\langle\left\langle \mathbf{x}_{n'}\delta_{tn,tn'}\right\rangle' \Big/ \left\langle\delta_{tn,tn'}\right\rangle'\right\rangle$ is $\mathbf{x}_k\,(t_n)$ averaged over all n, etc. The estimated position of the object center, is

$$\mathbf{xc}_k\,(t_n) = [\mathbf{x}_k\,(t_n) + \mathbf{R}_k\beta_k + \mathbf{V}_k t_n\beta_k]\,(\mathbf{1} + \beta_k)^{-1} \tag{7.2-28}$$

where $\beta_k = \mathbf{CX}_k^{-1}a_k^2$. This expression has an intuitive interpretation as a weighted average of two contributions (one, an estimated average position of pixels associated with object k in t_n-frame and another, an estimated track position in t_n-frame); and the weights are the inverse uncertainty measures associated with each piece of information: the size of object a_k^{-2} and the covariance of track $\mathbf{CX}_k^{-1}$.

7.2.9 Object-Track Declaration

Upon convergence of the iterative association–estimation procedure using any mixture of models described above, a decision should be made as to the actual presence of a target. This decision can be made by evaluating the target-to-clutter log-likelihood ratio (*LLR*) and comparing it to a threshold. A standard way of computing *LLR* is by estimating two models, a target model and a clutter one, computing their likelihoods, and taking the ratios. In our case containing multiple adaptive submodels (modes), this approach is not feasible. Instead, we compute a "local" *LLR* using the fact that each pixel has a certain pdf of being a target and being a clutter. In each frame we take the most likely target pixel and we compute

an *LLR* for this pixel. Then we average this number along the track. Here is the algorithm. First, compute the log-likelihood ratios for the k-target candidate pixel n, LLR_{nk},

$$LLR_{nk} = \log \text{ pdf } (\mathbf{z}_n | H_k) - \log \text{ pdf } (\mathbf{z}_n | H = \text{clutter}) \tag{7.2-29}$$

It is sufficient to consider few pixels on every frame around the candidate target position $\mathbf{X}_{kn}$. Second, select the maximum of LLR_{nk} on every frame,

$$LLR_{tn,k} = \max (LLR_{nk}) \text{ over } n \in t_n - \text{frame} \tag{7.2-30}$$

Third, compute the average log-likelihood ratio for the target-candidate k

$$LLR_k = \sum_{tn} LLR_{tn,k} \Big/ \sum_{tn} 1 \tag{7.2-31}$$

Fourth, if LLR_k exceeds the threshold, the candidate track is declared a target. The threshold selection is based on the operational requirements, such as the acceptable detection and false alarm rate.

7.3 Association and Tracking Using Shannon–Einsteinian MFT (SE-CAT)

Often, signal strength, or brightness of an image cannot be accurately modeled, whereas object motion models are more reliable. In those cases Shannon–Einsteinian MFT developed in Chapters 4 and 6 is more appropriate than the Bayesian MFT considered above. Remember that the Bayesian theory assumes that among considered adaptation models there is an exact model (in probabilistic sense), whereas Shannon–Einsteinian MFT does not make such an assumption and, instead, extracts as much information from the data as possible, using the given adaptive model. A neural implementation of concurrent association and tracking using Shannon–Einsteinian MFT is called SE-CAT (read: sea-cat).

7.3.1 Association and Tracking in Radar Spectral Data

In Chapter 6 we described the Einsteinian Neural Network (ENN) for modeling radar Doppler spectral data. Here, we modify ENN for concurrent association and tracking of objects in radar data. Typical data consist of the radar return signal S as a function of Doppler frequency ω, range R, azimuth θ, and time t. Doppler frequency is proportional to the radial velocity, v

$$v = -0.5c\omega / \Omega \tag{7.3-1}$$

where c is speed of light ($c \approx 3 \times 10^8$ m/s) and Ω is the frequency of the emitted radar signal. Most often we consider measurements as functions of time. It will be convenient to denote nth cell (pixel) measurements as

$$(S_n, \mathbf{z}_n), S_n = S(t_n), \quad \mathbf{z}_n = \mathbf{z}(t_n) = (r_n, \theta_n, v_n) \tag{7.3-2}$$

Shannon–Einsteinian adaptive similarity AZ-*LL* is given by

$$\text{AZ-}LL = \max_{\mathbf{S}} \left\{ \sum_n N_n \ln F(n) \right\} = \max_{\mathbf{S}} \left\{ \sum_n N_n \ln \sum_k r_k F(n|k) \right\} \tag{7.3-3}$$

Here $F(n|k)$ are conditional object models, $\mathbf{S}$ is a set of model parameters, N_n is a number of photons in cell n,

$$N_n = S_n/\hbar\omega'_n; \quad \omega'_n = \max\left(|\omega_n|, \Delta\omega\right) \tag{7.3-4}$$

and $\Delta\omega$ is a sampling interval ($\Delta\omega = \pi/T$, T is a time interval over which a single Doppler spectrum is computed, it is called coherent integration time; $\Delta\omega$ and T are determined by radar operations). Objects usually appear in these data as unresolved "blobs," therefore we model them by using Gaussian functions in range, azimuth, and Doppler velocity,

$$F(n|k) = \text{Norm } G\left(\mathbf{z}_n|k\right) = \text{Norm } G\left(\mathbf{z}_n|\mathbf{Z}_{kn}, \mathbf{C}_k\right) \tag{7.3-5}$$

Norm is a normalization constant defined so that $\Sigma_n \; F(n|k) = 1$ [compare to Chapter 6, e.g., Eq. (6.2-7)]. Gaussian densities are defined so that $\int G(x)dx = 1$, also $\int G(x)dx \approx \Sigma_x \; G(x)\Delta x$, therefore

$$\text{Norm } = (\Delta r \cdot \Delta\theta \cdot \Delta v) \tag{7.3-6}$$

Sampling intervals ΔR, $\Delta\theta$, and Δv are determined by the radar operational parameters.

The model of object motion is given by time dependence of $\mathbf{Z}_k\left(t_n\right)$. Radial velocity estimated from Doppler measurements is much more accurate than estimated azimuthal velocity. Therefore, we model object motion as second order in range and first order in azimuth:

$$\mathbf{Z}_k\left(t_n\right) = \left(R_k + V_k t_n + 0.5\, A_k t_n^2, \Theta_k + \dot{\Theta}_k t_n, V_k + A_k t_n\right) \tag{7.3-7}$$

The k-object azimuthal velocity model is denoted by $\dot{\Theta}_k$. Parameters of this model are $\mathbf{S}_k = \left(r_k, R_k, V_k, A_k, \Theta_k, \dot{\Theta}_k, \mathbf{C}_k\right)$. For simplicity, we consider a diagonal covariance matrix, $\mathbf{C}_k = \text{diag}\left(CR_k, C\Theta_k, CV_k\right)$. Combining these models with general Shannon's MFT [Eqs. (4.6-18) and (4.6-19)] we obtain dynamic equations of ENN CAT. (This is considered in Problem 7.3–1, following Sections 7.2-3 and 7.2-6.) Fuzzy object-track memberships are defined similarly to Chapter 6,

$$f(k|n) = r_k F(n|k)/F(n); \quad F(n) = \sum_k r_k F(n|k) \tag{7.3-8}$$

Angular brackets are defined as follows; for any function f_n,

$$\langle f_n \rangle = \sum_n N_n f(k|n) \; (f_n) \tag{7.3-9}$$

For R_k, V_k, and A_k we obtain a system of three linear equations for the parameters of each object-track:

$$\begin{aligned}
\langle r_n \rangle - R_k \langle 1 \rangle - V_k \langle t_n \rangle &= 0 \\
\langle r_n t_n \rangle + \alpha \langle v_n \rangle - R_k \langle t_n \rangle - V_k \left(\left\langle t_n^2 \right\rangle + \alpha \langle 1 \rangle \right) - A_k \left(0.5 \left\langle t_n^3 \right\rangle + \alpha \langle t_n \rangle \right) &= 0 \\
\left\langle r_n t_n^2 \right\rangle + \alpha \langle v_n t_n \rangle - R_k \left\langle t_n^2 \right\rangle - V_k \left(\left\langle t_n^3 \right\rangle + \alpha \langle t_n \rangle \right) - A_k \left(0.5 \left\langle t_n^4 \right\rangle + \alpha \left\langle t_n^2 \right\rangle \right) &= 0
\end{aligned} \tag{7.3-10}$$

Here, $\alpha = CR_k / CV_k$. Terms proportional to α are large. Equations for Θ_k, $\dot{\Theta}_k$ are similar to Section 7.2.2, Eq. (7.2-11):

$$\begin{aligned}
\langle \theta_n \rangle - \Theta_k \langle 1 \rangle - \dot{\Theta}_k \langle t_n \rangle &= 0 \\
\langle \theta_n t_n \rangle - \Theta_k \langle t_n \rangle - \dot{\Theta}_k \left\langle t_n^2 \right\rangle &= 0
\end{aligned} \tag{7.3-11}$$

And equations for other parameters, r_k and $\mathbf{C}_k$, are similar to those obtained previously in Chapter 6 or in Section 7.2.2:

$$\begin{aligned}
r_k &= N_k / N, \quad N_k = \langle 1 \rangle \\
\mathbf{C}_k &= \left\langle (\mathbf{z}_n - \mathbf{Z}_{kn}) (\mathbf{z}_n - \mathbf{Z}_{kn})^{\mathrm{T}} \right\rangle / \langle 1 \rangle
\end{aligned} \tag{7.3-12}$$

7.3.2 Association and Tracking of Spatiotemporal Patterns

Here we derive ENN CAT equations for estimation and tracking spatiotemporal patterns, such as hurricane and other weather patterns and whirly patterns. Consider image sequence data, such as satellite movies of whirly hurricane motion, with intensity $I_n = I(\mathbf{x}_n, t_n)$. Describing the dynamics of intensity $I(\mathbf{x}, t)$ from the first principles is an extremely complicated problem, however, local properties of image flow satisfy relatively simple laws related to the first principles. Image flow velocities, $\mathbf{v}_n$, often satisfy equations of the type

$$\mathbf{v}_n = \Phi_k (\mathbf{S}_k, \mathbf{x}_n) \tag{7.3-13}$$

Here, $\mathbf{S}_k$ are model parameters and Φ_k is a function determined by a physical law governing the image flow in the vicinity of the nth pixel in the image sequence. For example, a whirly motion of air or fluid in (x, y)-plane around $\mathbf{R}_k = (x_k, y_k, 0)$ can be described by

$$\Phi_k (\mathbf{S}_k, \mathbf{x}_n) = [\mathbf{c}_k \times \mathbf{R}_{kn}] / \mathbf{R}_{kn}^2 = c_k (-y_{kn}, x_{kn}) / \mathbf{R}_{kn}^2$$

where

$$\mathbf{R}_{kn} = \mathbf{x}_n - \mathbf{R}_k, \quad \mathbf{c}_k = (0, 0, c_k), \quad x_{kn} = x_n - x_k, \quad y_{kn} = y_n - y_k \tag{7.3-14}$$

To use Eq. (7.3-13), we have to estimate velocity $\mathbf{v}_n$ from the data. We consider several approaches to do this.

7.3.2.1 Simplified Image Flow Estimation

The image flow velocity can be estimated as follows. Assuming that in a small portion of an image all pixels move with the same velocity, $\mathbf{v}$, this local image flow is described by $I(\mathbf{x}, t) = I(\mathbf{x} + \mathbf{v}t)$, which satisfies the following equations:

$$I_t = \mathbf{v}\mathbf{I}_{\mathbf{x}}; \quad I_t = (\partial/\partial t) I(\mathbf{x} + \mathbf{v}t); \quad \mathbf{I}_{\mathbf{x}} = (\partial/\partial \mathbf{x}) I(\mathbf{x} + \mathbf{v}t) \tag{7.3-15}$$

The local image flow velocity $\mathbf{v}$ can be estimated from this equation using the least mean square procedure in the vicinity of $(\mathbf{x}_n, t_n)$; see Problem 7.3–2. The result is

$$\mathbf{v}_n = \left(\Sigma\, \mathbf{I}_x \mathbf{I}_x^{\mathrm{T}}\right)^{-1} (\Sigma\, I_t \mathbf{I}_x) \tag{7.3-16}$$

The sum here is taken over pixels in the vicinity of $(\mathbf{x}_n, t_n)$.

The Shannon–Einsteinian similarity for the model (7.3-13),

$$\text{AZ-}LL = \max_{\mathbf{S}} \left\{ \sum_n I_n \ln \sum_k r_k F(n|k) \right\}, \quad F(n|k) = G\left[\mathbf{v}_n | \Phi_k\left(\mathbf{S}_k, \mathbf{x}_n\right), \mathbf{C}_k\right] \tag{7.3-17}$$

can be maximized over the parameters $\mathbf{S}_k$ using the general MFT iterative Eq. (4.6-19),

$$\mathbf{S}_k = \mathbf{S}_k + \sum_n I_n f(k|n) [\partial \ln \Phi_k(\mathbf{S}_k, \mathbf{x}_n)/\partial \mathbf{S}_k\}, \quad f(k|n) = r_k F(n|k)/F(n) \tag{7.3-18}$$

7.3.2.2 General Image Flow Model

The least mean square estimation (7.3-16) of the image flow velocity Eq. (7.3-15) assumes a Gaussian density of the deviations from this equation. A general Gaussian mixture density of the deviations accounts for the multiple physical processes and noise that might be responsible for the image flow. Shannon–Einsteinian similarity for the multiprocess model of the type (7.3-15) is given by

$$\text{AZ-}LL = \max_{\mathbf{S}} \left\{ \sum_n I_n \ln \sum_k r_k F(n|k) \right\}, \quad F(n|k) = G\left(I_{tn} | \left(\mathbf{v}_k \mathbf{I}_{xn}\right), \mathbf{C}_k\right) \tag{7.3-19}$$

Solution of this problem results in a modification of (7.3-16):

$$\mathbf{v}_k = \left\langle \Sigma\, \mathbf{I}_{\mathbf{x}} \mathbf{I}_{\mathbf{x}}^{\mathrm{T}} \right\rangle^{-1} \langle \Sigma\, I_t \mathbf{I}_{\mathbf{x}} \rangle, \quad \text{where } \langle \ldots \rangle = \sum_n I_n\, f(k|n)(\ldots) \tag{7.3-20}$$

This approach gives the velocities of several "competing" or overlapping image flow processes. It can be used over a portion of an image, where all velocities $\mathbf{v}_k$ are constant.

7.3.3 CAT of Spatiotemporal Patterns Described by General PDE Models

In the general case, when image flow is determined by multiple processes, each controlled by a partial differential equation (PDE), the general ENN CAT approach is formulated as follows. Consider a k-process conditional PDE of the form

$$I_t(n) = \Phi_k\left(\mathbf{S}_k, \mathbf{x}_n, \partial/\partial \mathbf{x}_n\right) I(n) \tag{7.3-21}$$

Here, $I_t = (\partial/\partial t) I(n)$ is a temporal derivative and $\Phi_{kn} = \Phi_k\left(\mathbf{S}_k, \mathbf{x}_n, \partial/\partial \mathbf{x}_n\right)$ is a general spatial partial derivative operator. Shannon–Einsteinian similarity is given by

$$\text{AZ-}LL = \max_{\mathbf{S}} \left\{ \sum_n I_n \ln \sum_k r_k F(n|k) \right\}, \qquad F(n|k) = G\left[I_t(n)|\Phi_{kn} I(n), C_k\right] \tag{7.3-22}$$

This expression can be maximized over the parameters $\mathbf{S}_k$ using the general MFT Eq. (4.6-19). This gives the following iterative equation for the concurrent association and estimation of the parameters of the differential operator:

$$\begin{aligned} \mathbf{S}_k &= \mathbf{S}_k + \sum_n I_n f(k|n) \partial \left\{ \ln \left[r_k G\left[I_t(n)|\Phi_{kn} I(n), C_k\right]\right]\right\} / \partial \mathbf{S}_k \\ &= \mathbf{S}_k + \sum_n I_n f(k|n) \left[I_t(n) - \Phi_{kn} I(n)\right] / C_k \left[\partial \Phi_{kn} I(n) / \partial \mathbf{S}_k\right], \\ f(k|n) &= r_k F(n|k) / F(n) \end{aligned} \tag{7.3-23}$$

The ENN CAT method formulated here is suitable for complex situations, when image flow is determined by several spatiotemporal dynamic processes and each process is described by PD equations with unknown parameters. ENN CAT concurrently associates image pixels with processes by fuzzy variables $f(k|n)$ while estimating the unknown parameters of the processes.

7.3.4 Examples of Concurrent Association and Tracking in Radar Data

7.3.4.1 Example of Simulated Low-Signal Target

We illustrate application of the ENN CAT method using radar spectral data similar to those considered in Chapter 6. Figure 7.3-1 shows concurrent association and tracking of a low-signal target using OTH radar. One scan of data is shown in Fig. 7.3-1a; range is shown along the vertical axis; it contains 500 range bins (intervals) and covers 500 to 1400 km; the horizontal axis contains 128 Doppler velocity bins and covers an interval of ± 140 m/s. The main feature in data is the central clutter peak near zero Doppler, the dominating feature at all ranges: this is a ground/sea return. On each side of this peak there are nonzero Doppler returns due to ionospheric fluctuations and radar noise. There is about 100 dB between the ground peak and the noise floor, and there is a large number of smaller clutter peaks exceeding the noise floor by about 20 to 30 dB. A simulated target was inserted in this data in the upper-right corner moving with a velocity of about 90 m/s. The average strength of the target return is only about 5 dB (above the local noise level around the target) and it cannot be seen by the naked eye. The extent of the target return in range and Doppler is determined by the radar resolution (called ambiguity function) and is slightly larger than one range-Doppler cell. The target is moving with a constant velocity and its amplitude randomly fluctuates from scan to scan with a standard deviation of ± 6 dB (this is expected from a real target). Thus, the target returns are different from the model described in Section 7.3.1, which was used to track the target. The scan shown in Fig. 7.3-1 is one of 10 scans used for CAT.

The target cannot be detected on a single scan in Fig. 7.3-1a with any reliability. Therefore classical techniques that first detect targets and then estimate tracks are inapplicable.

It is necessary to perform concurrent association and tracking on multiple scans. Results of CAT processing are shown in Fig. 7.3-1b. This shows the estimated ENN CAT model that was described in Section 7.3.1, for the same scan as in Fig. 7.3-1a. Similar to Chapter 6, we used one uniform model for modeling the noise floor and two Gaussian models to model the central peak. These models are estimated for each of 500 range bins and each of 12 scans, so that there are 5000 clutter models estimated. In addition, we used four target models that were initiated symmetrically about the image with large covariances, so that initially, every cell in the image sequence had a nonzero probability to be a target cell. The target models are estimated from all 10 scans. On convergence, estimated models are shown in Fig. 7.3-1b for one of the scans. The declared target is shown in Fig. 7.3-1c. We used the declaration algorithm described in Section 7.2.9 with modified log likelihood computed according to Eq. (7.3-3). Only the target-type models are subject to a log-likelihood declaration test; the central ground peak in Fig. 7.3-1c is shown for reference purpose only.

A threshold value in the declaration algorithm is selected based on a tradeoff between detection probability and false alarm rate. Detailed procedures for threshold selection are not discussed here. Instead, a simple illustration of the issues involved is shown in Fig. 7.3-2. In this figure we show results of performing CAT for four different target signal strengths: we used target-to-clutter ratio values of 5, 8, 10, and 15 dB. The 5-dB case was illustrated in Fig. 7.3-1; the other three cases were different only in the strength of the target signal; clutter was exactly same. The vertical axis shows the number of occurrences of each of the log-likelihood values; in this example, it is trivial because we have only few data points: there are four different target log-likelihood values and three clutter values; each value occurred just once. All three clutter log-likelihood values are below 700 and all four target log-likelihood values are above 800, so selecting the threshold at 750 we rejected all clutter (zero false alarms) and detected all targets (100% detection).

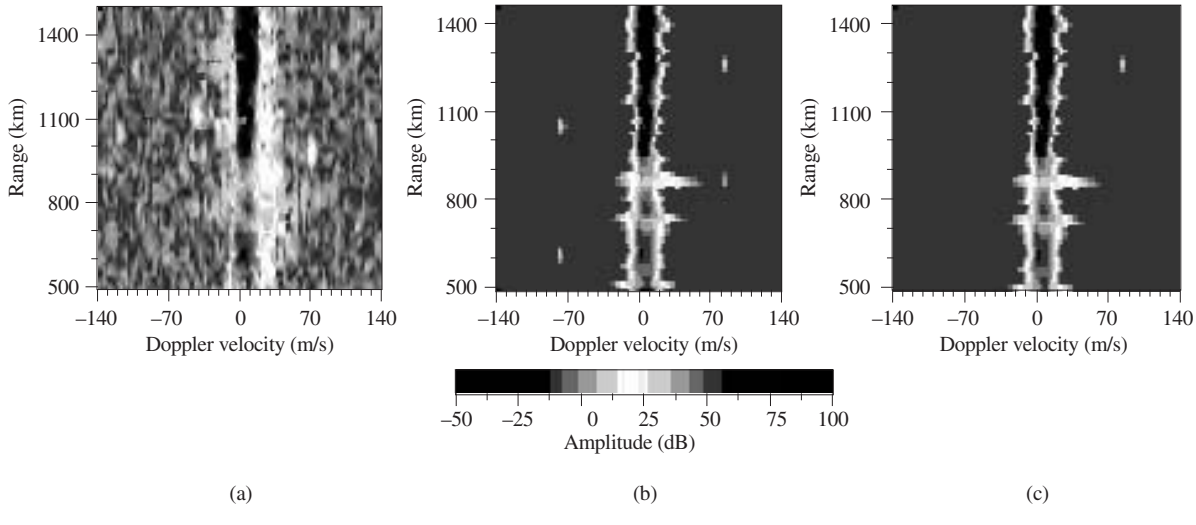


Figure 7.3-1 Concurrent association and tracking in OTH radar (t, ω, R) spectra; target signal-to-clutter ratio is 5 db; (**a**) data; (**b**) estimated model; (**c**) declared targets.

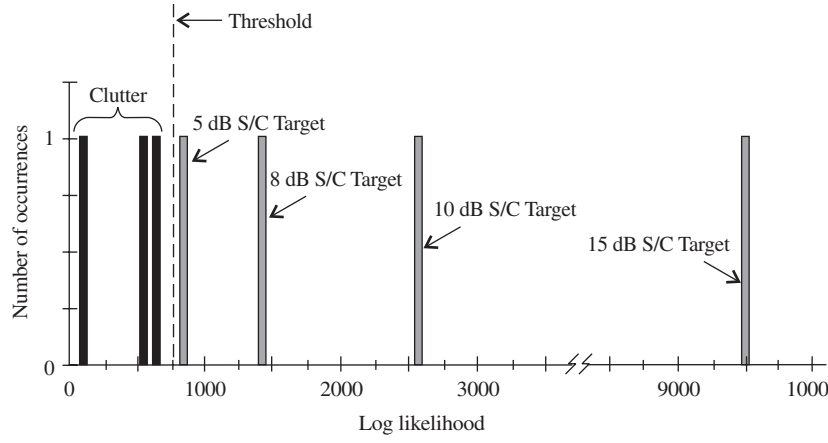


Figure 7.3-2 Log-likelihood values for clutter and targets; threshold is selected to yield 0 false alarms and 100% detection.

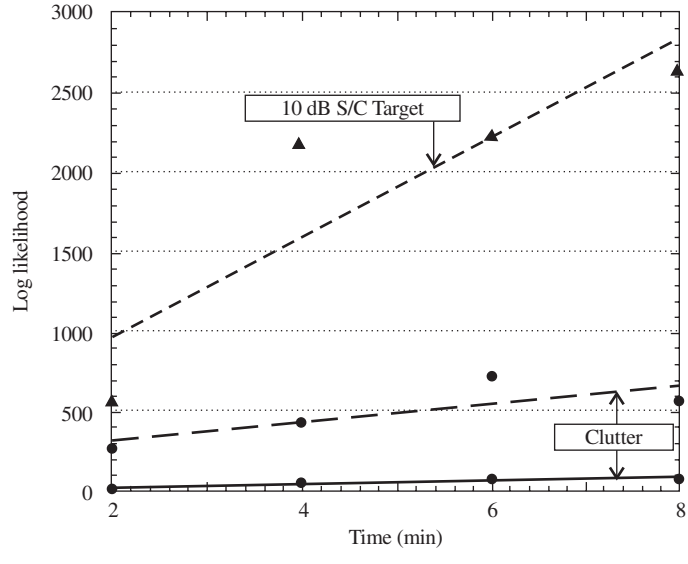


Figure 7.3-3 Cumulative log-likelihood values increase with time along the tracks; separability between the target and clutter increases as well.

The time dependence of the log-likelihood values is illustrated in Fig. 7.3-3 for the case of the 10-dB target. The vertical axis here shows cumulative log-likelihood values along the track, computed after 2, 4, 6, and 8 min of track (4, 8, 12, and 16 scans). For clutter, we show the maximum and minimum log-likelihood values at each time. Theoretically, the

target cumulative log-likelihood value is expected to be proportional to time, therefore, in addition to the data, we show fitted lines. As expected, separability between the target and clutter increases with the tracking time.

7.3.4.2 Example of Real Target with High Signal-to-Clutter Ratio

Figure 7.3-4a shows a similar data set containing a real target. This target signal is relatively high, about 25 to 30 dB, and it does not present a difficulty for any detection and tracking method. The target is clearly seen in the left part of an image. The estimated model is shown in Fig. 7.3-4b; the target likelihood is much higher than the clutter ones, so the declaration is easy. The details of the data and modeling are illustrated in Fig. 7.3-4c for a single range-bin spectrum containing the target.

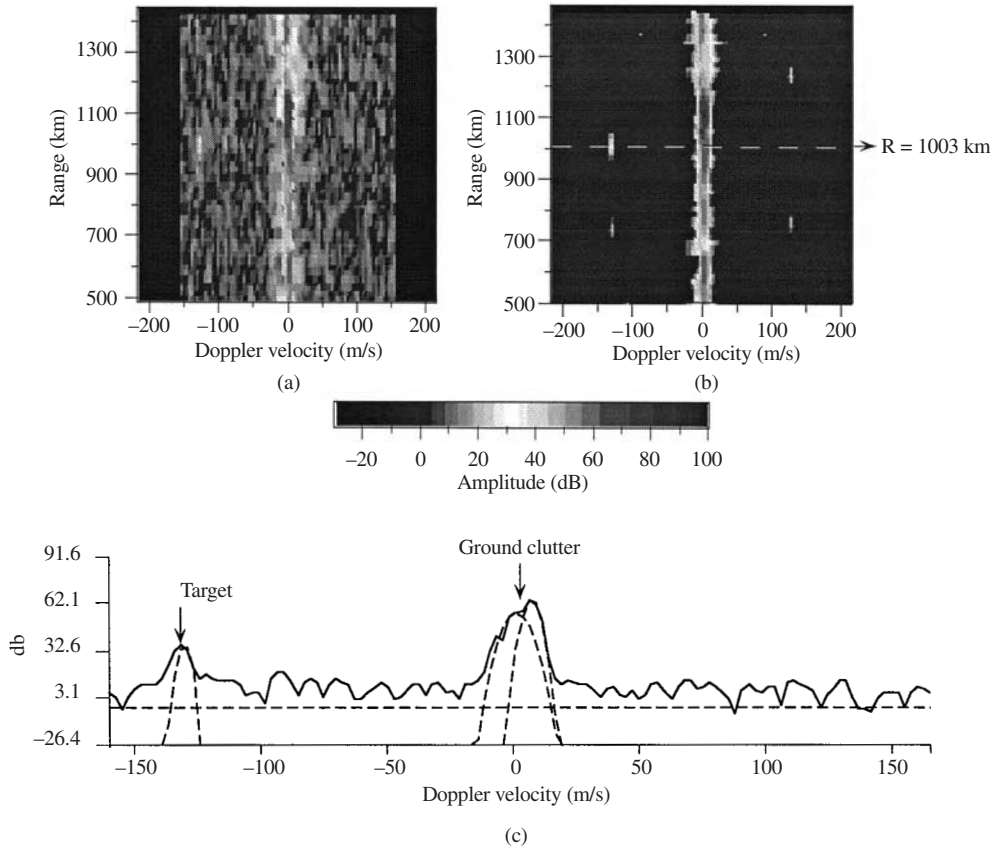


Figure 7.3-4 Cumulative association and tracking of a high signal-to-clutter target; (**a**) one scan of data; (**b**) estimated model of one scan; (**c**) estimated model of one range-bin spectrum containing target signal (at range $R = 1003$ km); the target was declared without false alarms (not shown).

7.4 Sensor Fusion MFT

An intelligent system has to be able to combine or "fuse" information from multiple sources, like sensors and databases. Fusion often improves performance; also training of adaptive systems and neural networks is enhanced by fusing teacher's information with self-learning. In this section MFT is extended to handle multiple information sources.

7.4.1 Information Fusion Problem

Fusion of information from multiple sources provides more information than a single source. Humans and animals routinely fuse data from multiple sources. Among applications of fusion technology are military as well as commercial problems, such as recognition of camouflaged targets using multiple sensors or "data mining" that is looking for useful patterns of information in multiple databases. A tremendous need for information fusion exists in computer networks, especially Internet and accessible from Internet databases. Fusion includes first, associating data from various sensors or databases with objects, second identifying objects using this information, and third, directing further searches based on available information and new incoming requests. This latter function is related to attention and it is considered in the next section. In this section we provide a unified mathematical formulation of a fusion problem and outline a general approach to extending MFT models considered in Chapters 5 and 7 to fusion.

7.4.2 Mathematical Formulation

Various sources of data or sensors are labeled using index $s = 1, \ldots, S$, and each data vector has this additional sensor index, $\mathbf{x}_{n,s}$. For different sensors, these vectors might have different dimensions and their components might have entirely different physical meanings (such as angle coordinates and intensity for a visual sensor or range and cross section for a radar). Also, the measurements by individual sensors are not necessarily synchronized in time or in space, so index n is a different index for each sensor, $n(s) = 1, \ldots, N(s)$; this complication of notations will be usually omitted unless required for clarity. Accordingly, a model under each hypothesis is a multisensor model that predicts observables for each of the sensors, $\mathbf{M}_{nks}(\mathbf{S}_k, t_n)$. Note that model parameters $\mathbf{S}_k$ describe the state of the object and do not necessarily depend on which sensor observes the object. With these changes, partial similarities $l(n, s|k)$ corresponding to alternative sources of data are given by pdf, $l(n, s|k) = r_k \text{ pdf}\left(\mathbf{x}_{n,s}|k\right)$, and the AZ-similarity, which is the total likelihood, is written in a manner similar to Eq. (7.2-6),

$$\text{LL} = \sum_{n;s} \ln \text{ pdf}\left(\mathbf{x}_{n,s}\right) = \sum_{n;s} \ln \left\{ \sum_{k} r_k \text{ pdf}\left(\mathbf{x}_{n,s}|k\right) \right\} \tag{7.4-1}$$

The only change here from the previous notations in this chapter is that index n is replaced with two indexes (n, s). Let us remember that in Section 7.2 we considered association of data with objects; the data were coming from multiple frames of a single sensor. Here, the data are coming from multiple frames of multiple sensors, or multiple databases. This

does not change the mathematical formulation except for an additional index s. All the association, tracking, estimation, and classification equations derived in Section 7.2 are directly applicable here. In particular, associations of data (indexed by n, s) and objects (indexed by k) are given by fuzzy memberships,

$$f(k|n,s) = P(k|n,s) = r_k \,\text{pdf}\left(\mathbf{x}_{n,s}|k\right) / \sum_{k'} r_{k'} \,\text{pdf}\left(\mathbf{x}_{n,s}|k'\right) \tag{7.4-2}$$

Again, the fuzzy memberships converge in the process of learning to the a posteriori Bayesian probabilities.

MFT fusion described here is a general method applicable to fusion problems of various types and complexities summarized in Table 2.10-1. For "simple" fusion problems, statistical models are used for defining pdf and probabilities. For complicated fusion problems, statistical models are augmented with dynamic or geometric models of the type considered in previous sections, or appropriate models should be developed as needed. For very complicated fusion problems, hierarchical models are needed that in addition to statistical, geometric, and dynamic information contain databases and knowledge bases representing the context, situations, or "entire world," within which the fusion is being performed.

7.4.3 Can Sensor Fusion Degrade Performance?

Making decisions with more information is better than otherwise. Still sensor fusion does not always guarantee improved performance (more data do not necessarily mean more information). In simple cases fusion always improves performance. These include cases in which data already have been associated with objects and when learning is not needed (so that pdfs can be estimated off line, e.g., because they do not change in real time). But in complicated adaptive cases, fusion could lead to degradation of performance due to two types of errors: association errors and classifier–estimation errors. The model-based approach to fusion provides a systematic foundation for association and fusion of data from any kind of diverse source. However, when adaptive models are used, the unknown model parameters have to be estimated from data. Estimation inherently involves errors, therefore the benefits of using additional information potentially contained in the data should be evaluated vs. a need for estimation of additional parameters and potential errors involved in the process. It is desirable to minimize the number of additional adaptive parameters used for fusion, especially when utilizing marginally useful data.

For example, when fusing based on the geographic location of the object, no additional parameters are needed to utilize additional sources of information: the only adaptive parameters are the coordinates of the object, the same parameters common for all sources. But when every source of information is characterized by its own set of adaptive parameters, the potential benefits have to be evaluated vs. potential error increase. This problem is not new for us: it is a particular case of the general problem of the model selection and a need to choose the optimal number of adaptive parameters.

Decisions about whether to use fusion or not, when based on experience, could be costly. Alternatively, they can be made with the help of simulations or by studying the mathematical performance bounds (Chapter 9), which account for the additional information in the data vs. a need for estimation of additional model parameters. Performance bounds or computer

simulation can be used to optimize fusion systems. Biological solutions have been optimized for billions of years. In biological systems there are "prewired" fusion levels. And, also, there are adaptive fusion decisions. When are they used? Likely, there are a priori models that guide decisions about adaptive fusion. Complex artificial systems also need these types of models. Their functional shapes can be determined by studying information-related performance bounds discussed in Chapter 9.

7.5 Attention

Here, we discuss a close relationship between information fusion and attention, and the role of attention is analyzed as a mechanism of efficient utilization of sensory resources in the process of adaptation and learning.

Attention directs limited sensory and processing resources to the objects of importance. In a hierarchical system, objects of importance are input signals at a given level. Note that unlimited resources, exceeding requirements of input signals, do not need a separate attention function. Criteria of importance depend on the goals of tasks being performed and usually multiple goals *and* multiple input signals compete for attention. The mathematics of this process has to account for diverse and competing criteria. Some of these are summarized in Table 7.5-1. At the top level in this table we differentiate two types of tasks: specific learning and general learning. Specific learning refers to the highest priority task at hand and is directed at maximizing certain specific values. General learning refers to updating and improving the internal model and acquiring knowledge (this includes maintaining the situational awareness). Within general learning, attention is subdivided among learning novel objects, discerning confusing objects, confirming well-recognized objects, and maintaining general awareness. Table 7.5-1 gives examples of attentional functions and identifies specific measures that can be used as attentional cues. Some of these cues originate within the same level in a hierarchy, and some are supplied by higher levels. Diverse functions and cues listed in Table 7.5-1 are combined by an intelligent system, e.g., by allocating a certain percentage of resources to each function.

TABLE 7.5-1
Attention Mechanisms and Cues

	Specific Learning, Task Related	General Learning			
Task		Recognition (Novel Object)	Confirmation (of Recognized Objects)	Resolution of Difficulties	Maintaining Situational Awareness
Cue/measure	Object-value (V) $f(k\|n)$-weighted $V^* f(k\|n) >$ th1	Probability of detection $f(k\|n) >$ th2	$f(k\|n) >$ th3	$f(k\|n) <$ th4 for all classes k	Random objects n
Origin of a cue in a hierarchy	Higher level (V), same level $f(k\|n)$	Same level			

For example, consider development of a sensor fusion system for detecting intruders. At the bottom level of a hierarchy several acoustic sensors for detecting novel objects could be used; these sensors operate continuously and require no attention mechanism. A simple internal model can be used with just two classes: noise and signal; and a low threshold th2 will ensure that all intruders will be detected at this level. When a signal exceeds the threshold, attention issues a request for an additional surveillance, and the next hierarchical level will issue a message directing a CCD camera toward the signal area. A relatively simple spatiotemporal internal model will be used to process a sequence of images, identify the size and motion speed of present objects, and estimate their class-membership functions. If any class-membership function $f(k|n)$ for potential-intruder class (k) exceeds the threshold (th3), a human operator might be called on. More levels of processing with more complex internal models can be used to reduce false alarms and enhance detection. A similar approach can be used for database searches. First, simple screening models can be employed to detect records of potential value. Then, more complex, multirecord models can be used to associate records among databases, etc. Activation of a concept by an intelligent agent (a submodel) also serves a role of an attention request to other agents, as illustrated in Fig. 7.5-1.

A specific application of fusion and attentional mechanisms is encountered in training neural networks and other learning algorithms and systems. Usually, developing training data is a time-consuming task. It could be significantly speeded up if supervised and unsupervised training is combined. In this application, a system first automatically clusters unlabeled objects into groups, then it requests a human teacher to provide labels (class names) for several well-recognized objects from each cluster $[f(k|n) > \text{th3}]$. Using this information, the objects are reclassified, and the procedure is repeated until there are no more errors among well-recognized objects. Then a system may ask a teacher for the class labels for the most confusing objects $[f(k|n) < \text{th4}]$; this will result in improving the details of the classifier boundaries. This kind of interactive operation of a human–machine system results in the gradual improvement of performance and fewer errors, while reducing demands for human intervention. Aided by such an intelligent system, a critical military

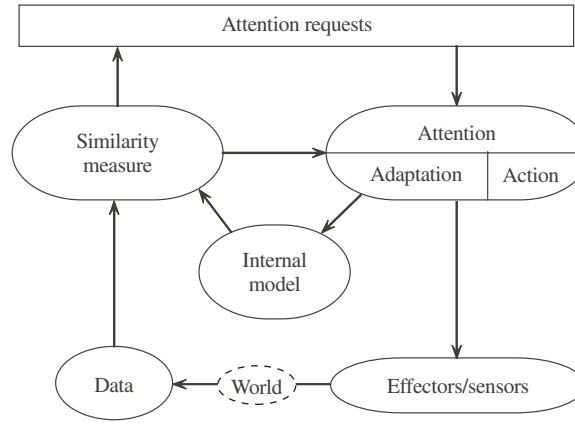


Figure 7.5-1 An MFT agent issuing and processing attention requests.

mission can be performed with reduced personnel. Or financial advisers can expand the diversity of the financial instruments they are tracking, while improving the timing of their decisions.

NOTES

1. If an object moves with a constant velocity in Cartesian coordinates, its velocity is not constant in angular coordinates of a typical visual camera. It is approximately constant, if an object is far away. Also, intensity is a function of range. If range can be estimated, a coordinate transformation to Cartesian coordinates can be performed, and a deterministic modeling of the intensity–range relationship can be used. Otherwise, a constant velocity and intensity model can be used as an approximation. Another approach is to estimate time-dependence adaptively. Decisions concerning fidelity and adaptivity of models ought to account for the available a priori knowledge, its reliability, and amount of data available for adaptation/estimation vs. the number of adaptive parameters (see also discussion in Section 7.4.3).
2. Deviations of the measured intensity from its mean value can be considered to be independent of deviations in object position, if these deviations are random, such as due to sensor errors or other random causes. However, when probabilistic densities are used to model nonrandom effects, this assumption is not necessarily valid. For example, flickering of object intensity can be caused by its jerky random movements about an approximately linear overall trajectory. In such a case, intensity and position variations are correlated and not independent.
3. When using a model with many links, a speed up of computations can be achieved by noticing that the system of Eqs. (7.2-19) is band limited. There are standard subroutines for fast solutions of band-limited systems.

BIBLIOGRAPHICAL NOTES

The contents of this chapter follow Perlovsky (1990b, 1991a, 1995), Perlovsky and Jaskolski (1994), and Perlovsky et al. (1995a,b, 1997a).

Data mining and revenue prediction example described in Section 7.1.5 (Muratet et al., 1998).

Additional reading on diverse neural attentional mechanisms (Grossberg, 1975, 1995; Grossberg and Schmajuk, 1987; Grossberg and Merrill, 1992).

PROBLEMS

7.1–1 Derive multivariable linear regression by estimating parameters of the linear model $\mathbf{y}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$. *Hints:* (1) Consider linear transformations, $\mathbf{x}' = \mathbf{L}_x(\mathbf{x} - \bar{\mathbf{x}}), \mathbf{y}' = \mathbf{L}_y(\mathbf{y} - \bar{\mathbf{y}})$. Define matrixes $\mathbf{L}_x$ and $\mathbf{L}_y$ so that $\mathbf{C}'_{xx} = \mathbf{L}_x\mathbf{C}_{xx}\mathbf{L}_x^{\mathrm{T}} = (1/N)\ \Sigma_n\ \mathbf{x}'\mathbf{x}'^{\mathrm{T}} = \mathbf{1}$, and $\mathbf{C}'_{yy} = \mathbf{L}_y\mathbf{C}_{yy}\mathbf{L}_y^{\mathrm{T}} = (1/N)\ \Sigma_n\ \mathbf{y}'\mathbf{y}'^{\mathrm{T}} = \mathbf{1}$, (here $\mathbf{1}$ stands for the diagonal unit matrix).

1.1. *Note:* this can always be accomplished by using the Gramm–Schmidt orthogonalization procedure, in which case matrixes $\mathbf{L}_x$ and $\mathbf{L}_y$ are called Choleski factors of the matrixes $\mathbf{C}_{xx}$ and $\mathbf{C}_{yy}$.

1.2. *Note:* $\mathbf{C}_{xx} = \mathbf{L}_x^{-1}\mathbf{L}_x^{-1\mathrm{T}}$, $\mathbf{C}_{yy} = \mathbf{L}_y^{-1}\mathbf{L}_y^{-1\mathrm{T}}$; $\mathbf{C}'_{yx} = (1/N)\ \Sigma_n\ \mathbf{y}'\mathbf{x}'^{\mathrm{T}} = \mathbf{L}_y^{-1}\mathbf{L}_x^{-1\mathrm{T}}$

2. Verify that $\left\{\min\ \Sigma_n \left[\mathbf{y}_n - \mathbf{A}\mathbf{x}_n - \mathbf{b}\right]^2\right\}$ is equivalent to $\left\{\min\ \Sigma_n \left[\mathbf{y}'_n - \mathbf{A}'\mathbf{x}'_n - \mathbf{b}'\right]^2\right\}$, where $\mathbf{A}' = \mathbf{L}_y\mathbf{A}\mathbf{L}_x^{-1}$ and $\mathbf{b}' = \mathbf{L}_y(\mathbf{b} - \overline{\mathbf{y}} + \mathbf{A}\overline{\mathbf{x}})$. The latter minimization is accomplished by solving two equations: $\partial/\partial\mathbf{A}'\{\ldots\} = 0$ and $\partial/\partial\mathbf{b}'\{\ldots\} = 0$. Show that $\mathbf{b}' = 0$, $\mathbf{A}' = \mathbf{C}'_{yx}\left(\mathbf{C}'_{xx}\right)^{-1}$. From this, derive the linear regression Eq. (7.1-6).

7.1–2 Derive multivariable linear regression as conditional expectation $\mathbf{y}(\mathbf{x}) = E\{\mathbf{y}|\mathbf{x}\}$, assuming the Gaussian joint pdf for $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, $\text{pdf}(\mathbf{z}) = G(\mathbf{z}|\mathbf{M}, \mathbf{C})$, where $\mathbf{M} = \left(\mathbf{M}_x, \mathbf{M}_y\right)$, and $\mathbf{C} = \begin{Bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{Bmatrix}$. *Hints:*

1. Perform a linear transformation as above, $(\mathbf{x}, \mathbf{y}) \rightarrow (\mathbf{x}', \mathbf{y}')$.
2. Perform a second linear transformation; define $\mathbf{y}'' = \mathbf{y}' - \mathbf{A}'\mathbf{x}'$ and $\mathbf{x}'' = \mathbf{x}'$ so that $\mathbf{C}''_{xy} = \mathbf{C}''_{yx} = 0 : E\left\{\mathbf{y}''\mathbf{x}''^{\mathrm{T}}\right\} = E\left\{\left(\mathbf{y}' - \mathbf{A}'\mathbf{x}'\right)\mathbf{x}'^{\mathrm{T}}\right\} = E\left\{\mathbf{y}'\mathbf{x}'^{\mathrm{T}}\right\} - \mathbf{A}'E\left\{\mathbf{x}'\mathbf{x}'^{\mathrm{T}}\right\} = \mathbf{C}'_{yx} - \mathbf{A}'\mathbf{C}'_{xx} = 0$. It follows that $\mathbf{A}' = \mathbf{C}'_{yx}\left(\mathbf{C}'_{xx}\right)^{-1}$.
3. Consider $\text{pdf}(\mathbf{y}'', \mathbf{x}'')$, since these variables are uncorrelated, $\text{pdf}(\mathbf{y}'', \mathbf{x}'') = \text{pdf}(\mathbf{y}'')\ \text{pdf}(\mathbf{x}'')$. It follows that $\text{pdf}(\mathbf{y}''|\mathbf{x}'') = \text{pdf}(\mathbf{y}'')$.
4. Prove that any linear transformation preserves the Gaussian shape of a pdf, while changing the means and covariances. In particular, substitute $\mathbf{y}''$ and $\mathbf{x}''$ into $G(\mathbf{z}|\mathbf{M}, \mathbf{C})$, and prove the 3 above. Show that the conditional mean of $\mathbf{y}''$, given $\mathbf{x}'$, $\mathbf{M}''_{y|x} = \mathbf{M}'_y - \mathbf{A}'\mathbf{x}'$. Transform back from $(\mathbf{x}'', \mathbf{y}'')$ to $(\mathbf{x}, \mathbf{y})$ and derive the regression equation.

7.1–3 Derive the estimation equation for multivariable linear autoregression, Eq. (7.1-15). Express all vector and matrix multiplications explicitly by using indexes. *Hints:*

1. Consider $\min\left\{\Sigma_t \left[\mathbf{x}_t - \Sigma_p\ \mathbf{A}_p\mathbf{x}_{t-p}\right]^2\right\}$;
2. Rewrite using indexes explicitly:

$$\left\{\sum_t \left[\mathbf{x}_t - \sum_p \mathbf{A}_p\mathbf{x}_{t-p}\right]^2\right\} = \left\{\sum_t\sum_i \left[x_t^i - \sum_{pj} A_p^{ij}x_{t-p}^j\right] \left[x_t^i - \sum_{p'j'} A_{p'}^{ij'}x_{t-p'}^{j'}\right]\right\}$$

3. Show that $\partial/\partial A_p^{ij}\{\ldots\} = 0$, leads to $\left\{\Sigma_t \left[x_t^i - \Sigma_{p'j'}\ A_{p'}^{ij'}x_{t-p'}^{j'}\right]\ x_{t-p}^j\right\} = 0$
4. Rewrite this as $C_p^{ij} - \Sigma_{p'j'}\ C_{pp'}^{jj'}A_{p'}^{ij'} = 0$
5. Show that index $\mathrm{l} = j + (p-1)\cdot D$ has a unique value for each $j = 1, \ldots, D$, $p = 1, \ldots, P$; and therefore in every instance, (j, p) can be substituted by l; obtain Eq. (7.1-15).

7.1–4 Apply the derived multidimensional autoregression to stock market prediction. Follow Chapter 1 Problems 1.3–7 through 1.3-11.

1. Add other indexes to your database, such as T-bills (or any other measure of interest rates), gold price, or foreign market indicators. Select an autoregressive model: which variables are you going to use (x), how many of them (D), and how long back

are you going to look (P). Estimate an autoregressive model and predict next day values from several previous day values. Determine how many past days you need for training/estimation: compute the number of parameters in all your $\mathbf{C}$-matrixes that you need to estimate from data; multiply by 10 (that many data points are needed to accurately estimate a single parameter) and divide by dimensionality (the number of variables you use every day: that many independent measurements); $N_{\text{par}} \sim P^* D^2/2$; days $= N^*_{\text{par}}\ 10/D \sim P^* D/2$.

2. Modify our autoregressive model by considering a prediction of a single variable (say DJ) from multiple variables. This changes the number of parameters and the required training interval. Did the result improve?
3. Play with your model, trying to improve it. Try different features/indicators. Optimize the training interval by verifying the stationarity assumption. Most mutual funds are happy when making a 15% profit per year (over several years). If you can do better, you are a pro! Even if you are doing much worse, do not get discouraged: professional traders are using leveraging techniques, which we did not discuss. Before investing your own money, make sure to try "paper trading" for few months and account for the brokerage fees. Continue reading the book.

7.1–5 Derive $P(m|\mathbf{x})$ in terms of the parameters of the joint density $\text{pdf}(\mathbf{x}, \mathbf{y})$. *Hints:*

1. Use the definitions, $P(m|\mathbf{x}) = \text{pdf}(\mathbf{x}|m)/\Sigma_{m'}\ r_{m'}\ \text{pdf}(\mathbf{x}|m')$, $\text{pdf}(\mathbf{x}|m) = \int \text{pdf}(\mathbf{x}, \mathbf{y}|m)\ d\mathbf{y}$;
2. Using results of Problem 7.1–2, show that for a Gaussian $\text{pdf}(\mathbf{x}, \mathbf{y}|m) \equiv G(\mathbf{x}, \mathbf{y}|\mathbf{M}_m, \mathbf{C}_m)$, with $(\mathbf{M}_m, \mathbf{C}_m)$ partitioned into $\mathbf{x}$ and $\mathbf{y}$ components as in Eq. (7.1-9),

$$\text{pdf}(\mathbf{x}|m) = G(\mathbf{x}|\mathbf{M}_{mx}, \mathbf{C}_{mxx})$$

3. It follows that

$$\text{pdf}(\mathbf{x}) = \sum_m r_m\, \text{pdf}(\mathbf{x}|m)$$

7.1–6 Apply the derived nonlinear regression to a next-day stock market prediction. Follow Problem 7.1–4. *Hint:* Identify $\mathbf{y}$ with $\mathbf{x}_t$ and $\mathbf{x}$ with $\mathbf{x}_{t-1}$.

7.1–7 Derive nonlinear autoregression estimation Eqs. (7.1-27) through (7.1-31). Combine derivations in Sections 5.6 and 7.1.2. *Hints:*

1. Write the log likelihood as

$$\text{LL} = \sum_t \ln\left\{\sum_m r_m\, G(\mathbf{x}_t, \ldots \mathbf{x}_{t-P}|\mathbf{M}_m, \mathbf{C}_m)\right\} =$$

$$\sum_t \ln\left\{\sum_m r_m\, G(\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-P}|\mathbf{M}_{mx}, \mathbf{C}_{mxx})\, G\left[\mathbf{x}_t|\mathbf{M}_{my}(\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-P}), \mathbf{C}'_{myy}\right]\right\}$$

where

$$\mathbf{M}_{my}(\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-P}) = \sum_p \mathbf{A}_{mp}\mathbf{x}_{t-p}$$

2. Take a derivative $\partial \text{LL}/\partial A^{ij}_{mp}$, as follows; first

$$\partial \mathrm{LL}/\partial A_{mp}^{ij} = \sum_t P\left(m|\mathbf{x}_t, \ldots, \mathbf{x}_{t-P}\right)$$

$$\partial/\partial A_{mp}^{ij} \ \ln G\left(\mathbf{x}_t|\mathbf{M}_{my}\left(\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-P}\right), \mathbf{C}'_{myy}\right)$$

second, compute the remaining derivative by using index notations, see Problem 7.1–3;

third, equate the result to 0, and multiply it by $\mathbf{C}'_{myy}$ (this will make the result independent of $\mathbf{C}'_{myy}$).

3. Compare the above equation with Problem 7.1–3 and derive Eqs. (7.1-31) for $\mathbf{a}_m^i$;
4. Derive Eqs. (7.1-31) for r_m and $\mathbf{C}_m$ similar to Section 5.6.

7.1–8 Consider modifications of the nonlinear autoregressive model for the case of $\mathbf{M}_{mx} \neq 0$. *Hints:*

1. Follow Section 5.6;
2. Show that $\mathbf{M}_{mx}$ is estimated in a similar way to the regular estimation of the mean in a Gaussian mixture model

$$\mathbf{M}_{mx} = \sum_t P\left(m|\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-P}\right) \ \mathbf{x}_t$$

3. Show that all covariances $\mathbf{C}_m$ are estimated in a similar way to the given equations by changing every $\mathbf{x}_t$ into $(\mathbf{x}_t - \mathbf{M}_{mx})$.

7.1–9 Derive Eq. (7.1-32) following the derivation of Eqs. (7.1-23) and (7.1-24).

7.1–10 Apply nonlinear autoregression to modeling stock market data.

7.2–1 Derive MLANS CAT estimation equations for the linear-track model. *Hints:*

1. Write the log likelihood $\mathrm{LL} = \Sigma_n \ \ln\{\Sigma_k \ r_k \ \mathrm{pdf}\ (\mathbf{z}_n|k)\}$;
2. Following Section 5.6, maximization of this log-likelihood expression can be achieved iteratively, by maximizing at each step $\Sigma_n \ P(k|n) \ \ln\{r_k \ \mathrm{pdf}\ (\mathbf{z}_n|k)\}$ over the parameters $\mathbf{S}_k$ of $\{r_k \mathrm{pdf}\ (\mathbf{z}_n|k)\} : \Sigma_n \ P(k|n) \ (\partial/\partial \mathbf{S}_k) \ \ln\{r_k \ \mathrm{pdf}\ (\mathbf{z}_n|k)\} = 0$
3. Substitute here Eqs. (7.2-7) and (7.2-8) and derive Eqs. (7.2-10) and (7.2-11).

7.2–2 Derive MLANS CAT estimation equations for the second-order track model. *Hints:* Repeat step 3 of Problem 7.2-1, replacing Eq. (7.2-13) instead of Eq. (7.2-7).

7.2–3 Derive MLANS CAT estimation equations for the link-track model. *Hints:* Repeat step 3 of Problem 7.2-1, replacing Eq. (7.2-16) with constraints (7.2-18) instead of Eq. (7.2-7). Use the method of Lagrange multipliers as described in Section 4.3-3.

7.2–4 Derive MLANS CAT estimation equations for the case of Doppler measurements. *Hints:*

1. Follow Problem 7.2–1.
2. Consider $\Sigma_n \ P(k|n) \ (\partial/\partial \mathbf{S}_k) \ \ln\{r_k \ \mathrm{pdf}\ (\mathbf{z}_n|k)\} = 0$, for $\mathbf{S}_k = (R_k, V_k)$
3. Evaluate the derivatives with respect to R_k and V_k:

$$(\partial/\partial R_k)\ \ln\{r_k\ \text{pdf}\ (\mathbf{z}_n|k)\} = (\partial/\partial R_k)\ \ln\{r_k\ G\left[r_n|\left(R_k + V_k\, t_n\right), CR_k\right]\}$$
$$= \left(r_n - R_k - V_k\, t_n\right)/CR_k$$
$$(\partial/\partial V_k)\ \ln\{r_k\ \text{pdf}\ (\mathbf{z}_n|k)\} = (\partial/\partial V_k)\ \ln\{r_k\ G\left[r_n|\left(R_k + V_k\, t_n\right), CR_k\right] G\left(v_n|V_k, CV_k\right)\}$$
$$= \left(r_n - R_k - V_k\, t_n\right) t_n/CR_k + \left(v_n - V_k\right)/CV_k$$

4. Combine 2 and 3:

$$\langle r_n\rangle - R_k\,\langle 1\rangle - V_k\,\langle t_n\rangle = 0$$
$$\left(\langle r_n t_n\rangle - R_k\,\langle t_n\rangle - V_k\left\langle t_n^2\right\rangle\right)/CR_k + \left(\langle v_n\rangle - V_k\,\langle 1\rangle\right)/CV_k = 0$$

5. From here, derive Eqs. (7.2-24).

7.2–5 Derive MLANS CAT estimation Eqs. (7.2-23) for tracking fuzzy circle. *Hints:*

1. Follow Problem 7.2–1.
2. Consider $\Sigma_{n'}\ P(k|n')\ (\partial/\partial \mathbf{S}_k)\ \ln\{r_k\ \text{pdf}\ (\mathbf{z}_{n'}|k)\} = 0$;
3. Evaluate the derivatives for $\mathbf{S}_k = [\mathbf{R}_k, \mathbf{V}_k, \mathbf{xc}_k\ (t_n)]$; account for $\partial \mathbf{xc}_k\ (t_{n'})\ /\partial \mathbf{xc}_k\ (t_n) = \delta_{tn,tn'}$;

$$(\partial/\partial \mathbf{R}_k)\ \ln\{r_k\ \text{pdf}\ (\mathbf{z}_{n'}|k)\} = (\partial/\partial \mathbf{R}_k)\ \ln\{G\left[\mathbf{xc}_k\ (t_{n'})\,|\mathbf{X}_{kn'}, \mathbf{CX}_k\right]\}$$
$$= \left[\mathbf{xc}_k\ (t_{n'}) - \mathbf{R}_k - \mathbf{V}_k t_{n'}\right]\mathbf{CX}_k^{-1}$$
$$(\partial/\partial \mathbf{V}_k)\ \ln\{r_k\ \text{pdf}\ (\mathbf{z}_{n'}|k)\} = (\partial/\partial \mathbf{R}_k)\ \ln\{G\left[\mathbf{xc}_k\ (t_{n'})\,|\mathbf{X}_{kn'}, \mathbf{CX}_k\right]\}$$
$$= \left[\mathbf{xc}_k\ (t_{n'}) - \mathbf{R}_k - \mathbf{V}_k t_{n'}\right] t_{n'}\mathbf{CX}_k^{-1}$$
$$\left[\partial/\partial \mathbf{xc}_k\ (t_n)\right]\ \ln\{r_k\ \text{pdf}\ (\mathbf{z}_{n'}|k)\} = \left[\partial/\partial \mathbf{xc}_k\ (t_n)\right]\ \ln\left\{G\left[\mathbf{x}_{n'}|\mathbf{xc}_k\ (t_{n'}), a_k^2\right]\right.$$
$$G\left[\mathbf{xc}_k\ (t_{n'})\,|\mathbf{X}_{kn'}, \mathbf{CX}_k\right]\} = \left\{\left[\mathbf{x}_{n'} - \mathbf{xc}_k\ (t_{n'})\right] a_k^{-2} - \left[\mathbf{xc}_k\ (t_{n'}) - \mathbf{R}_k - \mathbf{V}_k t_{n'}\right]\mathbf{CX}_k^{-1}\right\}\delta_{tn,tn'}$$

4. Substitute 3 into 2; from the last equation derive

$$\mathbf{xc}_k\ (t_n) = \left[\left\langle \mathbf{x}_{n'}\delta_{tn,tn'}\right\rangle' \Big/ \left\langle \delta_{tn,tn'}\right\rangle' + \mathbf{R}_k\beta_k + \mathbf{V}_k t_n \beta_k\right](\mathbf{1} + \beta_k)^{-1}$$

Here, $\beta_k = \mathbf{CX}_k^{-1} a_k^2$, and $\langle f_{n'}\rangle' = \Sigma_{n'}\ P(k|n')(f_{n'})$, for any $f_{n'}$.

5. Substitute $\mathbf{xc}_k\ (t_n)$ into the previous two equations for $\mathbf{R}_k$ and $\mathbf{V}_k$, and derive Eq. (7.2-26).

7.3–1 Derive ENN CAT estimation Eqs. (7.3-8) and (7.3-9). *Hints:*

1. Follow Problems 7.2–1 and 7.2-4. Maximization of the Shannon–Einsteinian AZ-*LL* can be achieved iteratively, by maximizing at each step $\Sigma_n\ N_n f(k|n)\ \ln\{A_k F(n|k)\}$ over the parameters $\mathbf{S}_k$ of $\{A_k F(n|k)\} : \Sigma_n\ N_n f(k|n)\ (\partial/\partial \mathbf{S}_k)\ \ln\{A_k F(n|k)\} = 0$;
2. Substitute here Eqs. (7.3-5) and (7.3-7) and derive Eqs. (7.3-8) and (7.3-9).

7.3–2 Derive image flow velocity estimation Eq. (7.3-16). *Hint:* Minimize the mean square difference between the left and right sides of Eq. (7.3-15) in the vicinity of $\mathbf{x} : \mathbf{min}_{\mathbf{v}}\left\{\Sigma\left[I_t - \mathbf{vI}_{\mathbf{x}}\right]^2\right\}$. The sum here is taken over pixels in the vicinity of $(\mathbf{x}_n, t_n)$.

chapter 8

QUANTUM MODELING FIELD THEORY (QMFT)

This chapter describes a quantum system implementing MFT. It is interesting from three standpoints, physical, engineering, and biological. From the point of view of physics, this chapter demonstrates that a quantum device can implement MFT. The engineering aspect is that quantum computers potentially offer a tremendous computational speedup. QMFT is a special purpose quantum computer. Even so it is not a general purpose computer, it implements a paradigm with wide application potential. The biological aspect is related to the possibility that microstructures of biological neurons are computational quantum devices, which play important role in neural information processing.

A note for the reader. The mathematical apparatus of quantum physics used is the familiar vector and matrix algebra. Peculiar notations used in physics are introduced as needed. This chapter does not contain any detailed exposition of quantum physics. For a reader not familiar with quantum physics, it might be difficult to comprehend relationships of the mathematical notations to physics. This chapter is not needed for an understanding of the rest of the book and could be skipped if difficult to understand.

8.1 Quantum Computing and Quantum Physics Notations

8.1.1 Quantum vs. Classical Computers

Quantum computing is a potential method of breaking through the limitations of the classical computing machines. It generated significant interest since Feynman draw attention to this area of research in 1972. The following two limitations of classical computational systems are expected to be surpassed. First, classical systems dissipate a finite amount of energy ($\sim kT$) per 1 bit for every operation. Second, a number of important problems in classical computational intelligence, in the number theory, and in other fields are very hard in that their solutions require a combinatorially large amount of computational steps as a function of the problem complexity. For example, a computation describing an electron transitioning from one state to another state requires taking integrals over all possible paths, which includes all combinations of path segments. Feynman turned this difficulty upside down: let the electron "compute" something of interest to us on every one of its paths. Then, in a single transition, this electron will accomplish a combinatorial number of computations.

We have discussed the conundrum of combinatorial complexity facing intelligent algorithms in Chapter 2; and classical MFT was designed as a solution to this problem. Quantum computing naturally performs "combinatorial speed-up" similar to MLANS. So let us combine the two concepts. It promises a tremendous speed-up in computations. Quantum systems do not dissipate energy (dissipation is a classical phenomenon). Quantum computation is expected to proceed without energy dissipation until the process of quantum measurement, which potentially can be postponed until the end of the computation process. And a quantum system can exist in a superposition of multiple states, so that multiple computational paths (including all possible combinations of path segments) can potentially be performed in parallel, within a single process of quantum interference between the quantum system states. In QMFT an internal dynamics of the model adaptation and association occurs as a process of quantum interference [in place of Eqs. (4.2-4) and (4.2-5)], and the final partition (segmentation, hypotheses choice, or classification, Ξ discussed in Section 4.1.3) is obtained in the process of quantum measurement.

8.1.2 Quantum Physics Notations and the QMF System

Let us introduce the necessary quantum physics notations and outline the main characteristics of QMFT. States of quantum systems are vectors in a Hilbert space. We use Dirac or bra-ket notations, in which states are denoted as "bra" $< \ldots |$ and "ket" $| \ldots >$. Ket is a vector and bra a transposed vector (and complex conjugated). QMFT describes a system that is characterized by quantum states $|k>$ and that interacts with the external world characterized by quantum states $|n>$. As in previous chapters, k numbers internal models and n numbers input data. The entire "universe" including the QMF system and the external world, in the general case, is described by a quantum state that is a superposition of states ($|k> |n>$),

$$|\Psi(t)> = \int_{N^*} \sum_{k=1}^{K} C_{kn}|k> |n> \tag{8.1-1}$$

Here integration over $n \in N^*$ includes spatial (and possibly other quantum coordinates of the external world) but excludes time, t. Quantum amplitudes C_{kn} are complex numbers related to fuzzy associations $f(k|n)$ according to the rules of quantum theory

$$f(k|n) = |C_{kn}|^2 \tag{8.1-2}$$

Generally, the QMF system is in a mixed state and is described by the density matrix,

$$\rho_{k1,k2}(t) = \int_{N^*} C_{k1,n}\, C^*_{k2,n} \tag{8.1-3}$$

or, equivalently, by the density operator

$$\rho(t) = \sum_{k1;k2=1}^{K} \left|k1> \rho_{k1,k2} < k2\right| \tag{8.1-4}$$

QMFT is designed to solve the same problem as solved by classical MFT. The states of the world $|n>$ correspond to the sensor data $\mathbf{X}_n$ and internal states of QMFT $|k>$ correspond to the concept-models $\mathbf{M}_k$. Consider a projection of the "universe" state (8.1-1) on the

specific world state $|n>$. This vector defines a state of the QMFT system corresponding to the world state $|n>$. We call it the encoding vector for the world state $|n>$, and denote it $|\mathbf{x}(n)>$,

$$|\mathbf{x}(n)>=<n|\Psi(t)>=\sum_{k=1}^{K} C_{kn}|k> \tag{8.1-5}$$

Identification of QMF and classical MF systems requires that fuzzy associations $f(k|n)$ have functional shape similar to that of the classical ones. This can be obtained as follows. Let us consider a wavefunction representation of the states of interest, $|k>$ and $|\mathbf{x}(n)>$. A wavefunction in quantum mechanics is an alternative description of states defined as a set of projections of a state vector on a basis vector set corresponding to some coordinate space. For example, if $\mathbf{z}$ is a usual 3-D coordinate space and $|z>$ is a Hilbert-space description of a state localized at point $\mathbf{z}$, the $\mathbf{z}$-space wavefunction of a state $|k>$ is given by

$$\psi_k(\mathbf{z}) = \langle \mathbf{z}|k\rangle \tag{8.1-6}$$

According to the definition, a wavefunction for the state localized at $\mathbf{z1}$, $|\mathbf{z1}>$, is a delta function,

$$\psi_k(\mathbf{z}) = \delta(\mathbf{z1}-\mathbf{z}) \tag{8.1-7}$$

Let us introduce a space of $\mathbf{j}$-coordinates, which we will call the encoding space. Coordinates $\mathbf{j}$ have the same dimension as our signal data $\mathbf{X}_n$ and models $\mathbf{M}_k$. (A multidimensional j-space can be realized by a multiparticle system.) The $\mathbf{j}$-space is defined so that the $\mathbf{j}$-space wavefunctions for the vectors of interest, $|k>$ and $|\mathbf{x}(n)>$, have a "natural" shape typical of many quantum-mechanical systems,

$$\psi_n(\mathbf{j}) = \langle j|\mathbf{x}(n)\rangle \sim \exp[i\mathbf{j}\mathbf{X}(n)]; \quad \psi_k(\mathbf{j}) = \langle j|k\rangle \sim \exp\left(i\mathbf{j}\mathbf{M}_k\right) \tag{8.1-8}$$

Note (1) that this shape of wavefunctions can occur only locally, around some values of $\mathbf{j}, \mathbf{X}(n), \mathbf{M}_k$, and (2) in most quantum-mechanical systems, wavefunctions have this local shape; and this is how we will understand these definitions, as approximate local relationships. Then,

$$C_{kn} = \int_{-\mathbf{J}}^{+\mathbf{J}} \psi_k^* \psi_n d\mathbf{j} \sim \int_{-\mathbf{J}}^{+\mathbf{J}} \exp\left[i\mathbf{j}\left(\mathbf{X}(n)-\mathbf{M}_k\right)\right]\, d\mathbf{j} \sim 2\ \mathrm{sinc}\left\{\mathbf{J}\left[\mathbf{X}(n)-\mathbf{M}_k\right]\right\} \tag{8.1-9}$$

Again, the integral here extends only over some locality where $\psi_n(\mathbf{j})$ and $\psi_k(\mathbf{j})$ overlap. A shape of the sinc-function here, when absolute value square is taken, is similar to the Gaussian functions used for conditional pdfs in Chapter 4. Note also that C_{kn} are normalized according to

$$\sum_{k=1}^{K} |C_{kn}|^2 = 1 \tag{8.1-10}$$

Let us analyze now the shape of $|C_{kn}|^2$ as a function of $[\mathbf{X}(n)-\mathbf{M}_k]$. Near $\mathbf{X}(n)=\mathbf{M}_k$, it is close to the Gaussian shape pdfs used in the numerator of $f(k|n)$ in Chapter 4, and globally,

it is normalized similar to $f(k|n)$. Thus, we conclude that with the above definition of the encoding states, the quantum probabilities, $|C_{kn}|^2$, have a functional form similar to that of classical MFT. This should not surprise us: a quantum system's fundamental property is that it is probabilistically distributed among several states, similar to the MF system, which attains the probabilistic distribution as a result of competition among alternative concept-models. If a specific parametric shape is important, it can be controlled by selecting appropriately shaped ψ-wavefunctions.

The proper encoding of the world state in terms of the QMFT states assumed above should be obtained as the result of the quantum dynamics of interaction between the QMF system and external world, the interaction that encodes the external information in the QMF system. The equations of motion of quantum dynamics are given by a Hamiltonian operator (a matrix in Hilbert space). Therefore, the next step is to define the Hamiltonian in such a way that the dynamics of a QMF system would lead to a solution of the same problem solved by classical MFT. Two types of quantum systems are considered below, first, a nonequilibrium quantum statistical system evolving to Gibbs distribution and second, a deterministic Hamiltonian quantum dynamic system.

8.2 Gibbs Quantum Modeling Field System

Here we define Hamiltonian so that the QMFT dynamics would lead to a Gibbs distribution with probabilities defined in a manner similar to fuzzy associations in (4.1-14) and (4.1-15) or to pdfs in (4.3-7) and (4.3-9). Following well-known principles of quantum statistical physics, we define the Hamiltonian through its relationship to the pdf. The dynamic variables of this system are unknown parameters $\{S_k^a\}$, while the data values $\mathbf{X}(n)$, $n = 1, \ldots, N$, are fixed, therefore

$$H = -\ln \text{ pdf}\left[\{S_k^a\}\,|\mathbf{X}(1), \ldots, \mathbf{X}(n)\right] \tag{8.2-1}$$

Using Bayes' theorem,

$$\begin{aligned}\text{pdf}\left[\{S_k^a\}\,|\mathbf{X}(1), \ldots, \mathbf{X}(N)\right] &= \text{pdf}\left[\{S_k^a\}, \mathbf{X}(1), \ldots, \mathbf{X}(N)\right] / \\ &\quad \text{pdf}\left[\mathbf{X}(1), \ldots, \mathbf{X}(N)\right] \\ &= \text{pdf}\left[\mathbf{X}(1), \ldots, \mathbf{X}(N)|\,\{S_k^a\}\right] \text{pdf}\left(\{S_k^a\}\right) / \\ &\quad \text{pdf}\left[\mathbf{X}(1), \ldots, \mathbf{X}(N)\right]\end{aligned} \tag{8.2-2}$$

Here pdf $\left(\{S_k^a\}\right)$ can be considered constant in the absence of prior information concerning values of these parameters. In this case, pdf $[\mathbf{X}(1), \ldots, \mathbf{X}(N)]$ is also constant, because in the absence of a priori values of these parameters it has to be invariant. Thus the Hamiltonian, Eq. (8.1-1), can be written as

$$H = -\ln \text{pdf}\left[\mathbf{X}(1), \ldots, \mathbf{X}(N)|\,\{S_k^a\}\right] + \text{const.} \tag{8.2-3}$$

Comparing Eq. (8.2-3) with Eqs. (4.3-7) and (4.3-9) and (4.2-8), we conclude that in the

absence of a priori information on the parameter values, the Hamiltonian is local and the Hamiltonian density H can be introduced

$$H = -\ln\{\text{pdf}[\mathbf{X}(n)]\}, \quad H = \int_{N(t)} H + \text{const.} \tag{8.2-4}$$

Here $N(t)$ stands for a set of observations available at time t. This emphasizes the fact that the Hamiltonian is defined using past data only (in case when n includes time). It is a nontrivial fact that availability of a priori values for the parameters may result in a nonlocal Hamiltonian.

The Hamiltonian Eq. (8.2-4) defines the dynamics of the density operator,

$$\rho(t) = \exp\left(-i\int_t H\right)\rho(0)\exp\left(i\int_t H\right) \tag{8.2-5}$$

To complete a correspondence between Gibbs QMF (GQMF) and classical MF, we define an operator of internal model parameters S as follows:

$$\mathbf{S} = \sum_{k\in K} |k> S_k^a <k| \tag{8.2-6}$$

Thus, the association–segmentation given by fuzzy associations $f(k|n)$ and the values of internal model parameters $\left\{S_k^a\right\}$ at time t can be obtained in the quantum measurement process,

$$f(k|n) = \text{Tr}[|k><k|\rho(t)], \quad S_k^a(t) = \text{Tr}[\mathbf{S}\rho(t)] \tag{8.2-7}$$

An initial state of a GQMF system is specified according to the initial state of the MF: one can choose initial values of S_k^a based on a priori phenomenological considerations; this leads to initial values $f(k|n)$ defined according to Eq. (4.2-4) and to initial values of the density matrix $\rho(0)$ defined according to Eq. (8.1-3) and (8.1-4),

$$\rho_{k1,k2}(0) = \int_{N^*} e^{-i\phi(k1|n)+i\phi(k2|n)}[f(k1|n)f(k2|n)]^{1/2} \tag{8.2-8}$$

The phases ϕ in this expression are left undefined and can be chosen to suit concrete problems at hand. One way to avoid a need to choose phases is to use an alternative initialization procedure: define a nonfuzzy initial association leading to a diagonal density matrix and compute initial values of the model parameters S_k^a from Eq. (8.2-6).

In a GQMF system described above, the finite temperature of the system (and, therefore, the finite accuracy of the computation) is an essential part of the system dynamics. On the one hand, this is a highly desirable property for any practical implementation of a quantum computing system. On the other, interaction with a thermal reservoir leads to irreversible energy dissipation processes involving quantum measurements. A desirable compromise is to reduce interactions with a thermal reservoir to a relatively rare occasion, which will ensure Gibbs distribution and will correct accumulating phase errors, while in between these interactions a GQMF system will evolve according to the Schrödinger equation (8.2-5) without energy dissipation.

8.3 Hamiltonian Quantum Modeling Field System

We define the Hamiltonian so that the Hamiltonian–QMF (HQMF) dynamics leads to a solution of the considered MFT problem. We will do this for a simplified case of MF, when the model values $\mathbf{M}_k$ *are* the model parameters, and we will consider all the covariances $\mathbf{C}_k$ equal to 1. In this case, the ML estimation equations for the parameters $\mathbf{M}_k$ are given by (5.2-8)

$$\mathbf{M}_k = \sum_n f(k|n)\mathbf{X}(n) / \sum_n f(k|n) \tag{8.3-1}$$

where the denominator has the meaning of the average number of observations classified to hypothesis k,

$$N_k = \sum_n f(k|n) \tag{8.3-2}$$

Correspondingly, in place of Eq. (8.2-6) we have

$$\mathbf{M} = \sum_{k=1}^{K} |k > \mathbf{M}_k < k| \tag{8.3-3}$$

An internal HQMF encoding $|\mathbf{x}(n) >$ of the external patterns $\mathbf{X}(n)$ (equivalently, of the states $|n >$ of the external world) is defined according to (8.1-5). Combining (8.1-3), (8.1-4), and (8.1-5),

$$\rho(t) = \sum_n |\mathbf{x}(n) >< \mathbf{x}(n)| \tag{8.3-4}$$

Let us also introduce an observation operator $\mathbf{X}$,

$$\mathcal{X} = \sum_{n=1} |\mathbf{x}(n) > \mathbf{X}(n) < \mathbf{x}(n)| \tag{8.3-5}$$

Consider eigenstates $|\lambda >$ of this operator,

$$\mathcal{X}|\lambda > = \lambda|\lambda >, \quad \text{or} \quad < \lambda|\mathcal{X}|\lambda > = \lambda \tag{8.3-6}$$

Substituting Eq. (8.3-5) into Eq. (8.3-6),

$$\lambda = \sum_{n=1} | < \lambda|\mathbf{x}(n) > |^2 \, \mathbf{X}(n) \tag{8.3-7}$$

Comparing this to Eqs. (8.1-2), (8.3-1), (8.3-2), and (8.3-3), we see that λ is identified with $\mathbf{M}_k \cdot N_k$, and $|\lambda >$ is identified with $|k >$. Introducing an operator

$$\mathcal{M} = \sum_{k \in K} |k > \mathbf{M}_k \cdot N_k < k| \tag{8.3-8}$$

we have the following identity for the quantum operators

$$\mathcal{M} = \mathcal{X} \tag{8.3-9}$$

This identity should be attained in the result of an internal QMF dynamics. States $|k>$, which are the eigenstates of the operator $\mathcal{M}$, should evolve from their initial states (according to the Schrödinger equation) into the eigenstates of $\mathcal{X}$, and the considered problem of MFT is defined as the problem of finding eigenstates of the operator $\mathcal{X}$, or, more accurately, finding a procedure that will evolve initial states into the eigenstates of $\mathcal{X}$. A number of algorithms exist that can be used for this purpose. These algorithms utilize unitary transformations and can serve as a basis for the design of a quantum system. Let us define the Hamiltonian H as

$$H(t) = i[\mathcal{M}, \mathcal{X}] \tag{8.3-10}$$

This Hamiltonian defines a dynamics that evolves eigenstates $|k>$ of an operator $\mathcal{M}$ into the eigenstates of operator $\mathcal{X}$. It might be noted that a dynamic diagonalization of an operator naturally occurs in many quantum systems, therefore we have a wide choice of a specific physical realization that will determine the specific realization of the Hamiltonian.

If $\mathbf{X}(n)$ are vector quantities (as they usually are), then $|\mathbf{x}(n)>$ are defined with a corresponding vector index, so that $\mathcal{X}$ is a vector operator and various components of $\mathcal{X}$ operate on the corresponding components of $|\mathbf{x}(n)>$. Thus components of $\mathcal{X}$ commute (and similarly, components of $\mathcal{M}$ commute). If values of N_k are known, then $\mathbf{M}_k$ can be directly obtained from $\mathcal{M}$. When N_k values are not known a priori, Eq. (8.3-2) cannot serve as a definition of the operator N, and N_k have to be obtained similarly to $\mathbf{M}_k$, in the process of internal HQMF dynamics. By comparing Eqs. (8.1-2), (8.1-3), (8.1-4), (8.1-5), (8.3-2), and (8.3-4), the N operator is identified with the density operator. Its eigenstates are different from $|k>$, N and $\mathcal{M}$ do not commute, and expected values of N in $\mathcal{M}$-eigenstates $|k>$ are given by the diagonal elements of the density matrix,

$$N_k = <k|\rho|k> \tag{8.3-11}$$

The HQMF system defined above evolves according to Schrödinger (Hamiltonian) dynamics without energy dissipation. We would also note that this algorithm does not require an exponential number of interfering quantum states; an exponential "speed-up" relative to the classical computation occurs in the result of interference between a number of states, which is only a linear function of the complexity of the system. Intuitively, it seems, however, that a required coherency (accuracy of computation) is a constant number that does not grow with the complexity of the problem. This is because, for every state $|\mathbf{x}(n)>$, an accurate computation is required of only few amplitudes $<k|\mathbf{x}(n)>$ with the highest probabilities $|<k|\mathbf{x}(n)>|^2$. In addition, the HQMF system has the advantage of a relatively simple formulation, providing a foundation for a physical realization of an MFT quantum computer.

BIBLIOGRAPHICAL NOTES

Original publications on QMFT (Perlovsky, 1997c,f).

Original Feynman papers on quantum computing (1982, 1986); these papers also contain references to the preceding work.

A contemporary review of the principles of quantum theory (Sakurai, 1985).

The density matrix describing quantum states was introduced in Neumann (1927); it is reviewed in Feynman (1972) and Sakurai (1985).

Principles of quantum statistical physics (Feynman, 1972; Zubarev, 1971).

Algorithms that evolve initial states into the eigenstates of X (Brockett, 1991; Oja, 1992; Xu and Yuille, 1995). The Hamiltonian (8.3-10) was introduced in Brockett (1991).

Relationship between quantum theory and statistical pattern recognition leading to equations of the type (8.3-11) was developed in Garvin and Perlovsky (1995).

PROBLEM

Design a quantum device implemeting QMFT. (This is a complicated problem good for a doctoral thesis. Its solution will lead to a breakthrough in computational devices.)

chapter 9

FUNDAMENTAL LIMITATIONS ON LEARNING

This chapter is concerned with the fundamental, mathematical limit on speed of learning. It is related to the limited amount of information present in data and a priori models, and it is measured by how many examples are needed to learn model parameters (e.g., a classifier or track parameters). This limit is established by the Cramer–Rao theorem, which is a fundamental result of mathematical statistics. The chapter discusses measures of speed of learning, the theoretical basis for the limit, and relationships between the theoretical limit and MLANS performance. At the end, I briefly discuss the possibility of extending the Cramer–Rao theory to the entire evolution and learning process of any population of intelligent systems.

9.1 The Cramer–Rao Bound on Speed of Learning

The Cramer–Rao bound (CRB) establishes fundamental mathematical limits on speed of learning. These limits are inherent to the available information in the data and models, and are independent of the approach to using that information, based on algorithms or neural networks. The CRB has been used for a long time in signal processing and other relatively simple estimation problems, such as tracking a single target without noise or clutter. In addition to discussing some of these classical CRB, this chapter describes CRB for more complicated problems, such as classification, or tracking multiple targets, or tracking in clutter, involving segmentation of data among alternatives.

9.1.1 CRB, Neural Networks, and Learning

Adaptivity and learning are primary features of neural networks. It is not surprising therefore that the problem of fast learning is among the most important problems in the area of neural networks. But what is a fast learning? How one can quantify and compare learning in different neural paradigms applied to diverse problems? The CRB provides a theoretical framework for the analysis of this problem.

The Cramer–Rao (CR) theorem establishes lower bounds for the statistical errors of estimated quantities, irregardless of the estimation process. In conventional nonparametric neural networks the weights are the quantities that are estimated from the data. In MFT, model parameters are the estimated quantities. In tracking a single target, the accuracy of the

track parameters is of primary interest and CRB are widely used in tracking applications. But when tracking in clutter, the ability to associate data and to establish track is even more important. In classification, the accuracy of classification is of primary interest as a measure of performance. In prediction, especially financial market prediction, timing is often more important than other characteristics. Although performance characteristics of primary interest are not necessarily the model parameter values, errors in model parameters can be related to other performance measures. The CRB accounts for the probabilistic uncertainty and does not directly address other issues such as fuzzy uncertainties of deterministic models related to the approximate nature of models and the related issue of robustness. Notwithstanding this limitation, the CRB is a powerful tool that helps quantify and compare performances of various techniques. It is also used for the diagnostics of existing algorithms to identify research areas in which algorithmic improvements are possible. We first derive CRB for model parameters, and then relate CRB to the performance characteristics of interests in specific applications. Several well-known classical CRB results are discussed as well as recently derived CRB for more complex MLANS models involving concurrent association/segmentation and parameter estimation.

9.1.2 Classical CRB for the Gaussian Means

A well-known example of a CRB is an accuracy of average value. Consider estimation of a one-dimensional mean M of a Gaussian distribution from N one-dimensional measurements x_n; all x_n are coming from a single class, which has a Gaussian distribution with the standard deviation σ. Sometimes we will denote σ as $\sigma\{x\}$, to show explicitly that σ is the standard deviation of the random variable x. In this chapter we differentiate between the true value of a parameter and its estimated value, by putting a ˆ (hat) over the estimated values. An often used estimate of M, $\hat{M}$ (read M-hat) is given by an average value,

$$\hat{M} = \sum_n x_n / N \tag{9.1-1}$$

The statistical variance of the average value is well known to be

$$\text{var}\{\hat{M}\} = E\left\{(\hat{M} - M)^2\right\} = \sigma^2/N \quad \text{or} \quad \sigma\{\hat{M}\} = \sqrt{\text{var}\{\hat{M}\}} = \sigma/\sqrt{N} \tag{9.1-2}$$

This value turns out to be the CRB for the estimation of the mean M; we will prove it in the following sections. A neural network or algorithm implementing this estimation is called an efficiently learning one, because no other approach can yield a more accurate mean value with the same number N of learning samples. In a multidimensional case, the CRB for estimating the mean $\mathbf{M}$ of a multidimensional Gaussian distribution $G(\mathbf{x}|\mathbf{M}, \mathbf{C})$, is given by

$$\text{cov}\{\hat{\mathbf{M}}, \hat{\mathbf{M}}\} = \mathbf{C}/N \tag{9.1-3}$$

Geometrically, one can think about this bound as an ellipsoid in $\mathbf{x}$-space given by

$$E\left\{(\hat{\mathbf{M}} - \mathbf{M})^{\mathrm{T}}\mathbf{C}^{-1}(\hat{\mathbf{M}} - \mathbf{M})\right\} = 1/N \tag{9.1-4}$$

The shape of this CRB ellipsoid is determined by the covariance matrix $\mathbf{C}$, and its size is determined also by the number of objects N.

An estimation of the mean given by the average value attains the CRB if the covariance matrix is known. If the covariance matrix is unknown and is to be estimated from the data, no estimation process can reach the accuracy of the CRB even in this relatively simple case. However, the standard average estimation above is asymptotically efficient, that is, it tends quickly to the CRB with increasing number of observations N.

9.1.3 CR Theorem

An accuracy of an estimated scalar parameter $\hat{S}$ is conventionally measured by its standard deviation $\sigma\{\hat{S}\}$, or by variance that is the square of the standard deviation $\text{var}\{\hat{S}\} = \sigma^2\{\hat{S}\}$. In case of a set of parameters, $\mathbf{S} = \{S^a, a = 1, \ldots, A\}$, its estimation accuracy is measured by the covariance matrix, $\text{cov}\{\hat{S}^a, \hat{S}^b\}$. This discussion implies that we observe a random quantity (or a set of random quantities) $\mathbf{x}$, and there is a $\text{pdf}(\mathbf{x})$ that depends on a set of parameters $\mathbf{S}$. The values of $\mathbf{S}$ are unknown and we consider estimate $\hat{\mathbf{S}}(\mathbf{x})$ to be a function depending only on $\mathbf{x}$ (and not on unknown $\mathbf{S}$). An estimate approximates $\mathbf{S}$ in some way. Parameters $\mathbf{S}$ are nonrandom, but their estimate $\hat{\mathbf{S}}(\mathbf{x})$ is random, since it is a function of random variables $\mathbf{x}$. Now we obtain a general expression for the CRB, which gives the lower bound for an average error in $\hat{\mathbf{S}}$.

Definition. Estimation is called unbiased if its expected value equals the exact value of the parameter

$$E\{\hat{\mathbf{S}}\} = \mathbf{S} \quad \text{or} \quad \int \hat{\mathbf{S}}\, \text{pdf}(\mathbf{x})\, d\mathbf{x} = \mathbf{S} \tag{9.1-5}$$

Let us take a derivative of both sides of this expression with respect to $\mathbf{S}$. Since, $\hat{\mathbf{S}}$ is not a function of $\mathbf{S}$,

$$\int \hat{\mathbf{S}}\, (\partial/\partial\mathbf{S})\, \text{pdf}(\mathbf{x})\, d\mathbf{x} = 1 \tag{9.1-6}$$

This can be rewritten as

$$\begin{aligned} 1 &= \int \hat{\mathbf{S}}\, [(\partial/\partial\mathbf{S})\ \ln\ \text{pdf}(\mathbf{x})]\, \text{pdf}(\mathbf{x})\, d\mathbf{x} = E\{\hat{\mathbf{S}}[(\partial/\partial\mathbf{S})\ \ln\ \text{pdf}(\mathbf{x})]\} \\ &= E\left\{\hat{\mathbf{S}} \cdot (LL_{;\mathbf{S}})^{\mathrm{T}}\right\} \end{aligned} \tag{9.1-7}$$

Here, $LL = \ln\ \text{pdf}(\mathbf{x})$; $(\partial/\partial\mathbf{S})$ is denoted using a $(;\mathbf{S})$ subscript,

$$(\partial/\partial\mathbf{S})\ \ln\ \text{pdf}(\mathbf{x}) = (LL_{;S})^{\mathrm{T}} \tag{9.1-8}$$

and superscript T denotes a transposed vector [$(\partial/\partial\mathbf{S})$ is a transposed vector, this can be verified by using indexes, and it is shown later in Problem 9.1-3]. According to the pdf normalization, $\int \text{pdf}(\mathbf{x})\, d\mathbf{x} = 1$, so

$$(\partial/\partial\mathbf{S}) \int \text{pdf}(\mathbf{x})\, d\mathbf{x} = 0$$

This can be rewritten as

$$(\partial/\partial\mathbf{S})\int \text{pdf}(\mathbf{x})\,d\mathbf{x} = \int (\partial/\partial\mathbf{S})\ \text{pdf}(\mathbf{x})\,d\mathbf{x} = \int [(\partial/\partial\mathbf{S})\ \ln\ \text{pdf}(\mathbf{x})]\,\text{pdf}(\mathbf{x})\,d\mathbf{x} = 0 \tag{9.1-9}$$

or

$$E\left\{(LL_{;\mathbf{S}})^{\mathrm{T}}\right\} = E\{LL_{;\mathbf{S}}\} = 0 \tag{9.1-10}$$

Combining this with (9.1-7), we obtain (see Problem 9.1–1),

$$\text{cov}\left\{\hat{\mathbf{S}}, (LL_{;\mathbf{S}})^{\mathrm{T}}\right\} = 1 \tag{9.1-11}$$

A product of two variances is no smaller than the covariance square, $\text{var}\{a\}\cdot\text{var}\{b\} \geq \text{cov}\{a, b\}$, therefore

$$\text{var}\{\hat{S}\}\cdot\text{var}\{LL_{;\mathbf{S}}\} \geq 1 \tag{9.1-12}$$

This gives the Cramer–Rao bound for the lowest value of expected error $\sqrt{\text{var}\{\hat{\mathbf{S}}\}}$,

$$\text{var}\{\hat{\mathbf{S}}\} \geq \mathbf{IM}^{-1}, \quad \mathbf{IM} = \text{var}\{LL_{;\mathbf{S}}\} \tag{9.1-13}$$

where **IM** is called an information matrix. Using indexes, this can be written in the component form,[1]

$$IM^{ab} = \text{cov}\{LL_{;a}, LL_{;b}\}, \quad \text{cov}\left\{\hat{S}^a, \hat{S}^b\right\} \geq \left(\mathbf{IM}^{-1}\right)^{ab} \tag{9.1-14}$$

Consider a set of statistically independent observations, $\{\mathbf{x}_n, n = 1, \ldots, N\}$. Its log likelihood is given by a sum of individual loglikelihoods:

$$LL = \sum_n LL_n = \sum_n \ln\ \text{pdf}(\mathbf{x}_n) \tag{9.1-15}$$

Therefore, when $\hat{\mathbf{S}}$ is estimated from a set of independent observations, $\{\mathbf{x}_n\}$, see Problem 9.1–2,

$$\mathbf{IM} = \text{var}\{LL_{;\mathbf{S}}\} = \text{var}\left\{\sum_n LL_{n;\mathbf{S}}\right\} = N\ \text{var}\left\{LL_{1;\mathbf{S}}\right\} = N\,\mathbf{IM}_1 \tag{9.1-16}$$

where LL_1 and $\mathbf{IM}_1$ are a log-likelihood and information matrix for an individual observation, and

$$\text{var}\{\hat{\mathbf{S}}\} = \text{var}_1\{\hat{\mathbf{S}}\}/N \geq \mathbf{IM}_1^{-1}/N \tag{9.1-17}$$

Let us emphasize that all parameter values in CRB expressions [(9.1-14), (9.1-15), (9.1-17)] are the true (not estimated) quantities. A numerical evaluation of a CRB involves taking expected values (integration) over only D dimensions of $\mathbf{x}$, even when estimation is performed using $D\cdot N$ scalar measurements contained in $\{\mathbf{x}_n, n = 1, \ldots, N\}$.

For the estimation of the mean of a Gaussian distribution, $\mathbf{S} = \mathbf{M}$. Evaluation of $\mathbf{IM}_1$ yields (see Problem 9.1–3)

$$\mathbf{IM}_1 = \mathrm{var}\{LL_{1;\mathbf{M}}\} = \mathbf{C}^{-1} \tag{9.1-18}$$

This leads to the CRB expression for the mean, given in the previous section Eq. (9.1-3).

Consider perfectly supervised learning with Gaussian class-conditional distributions. Perfect supervision assigns every pixel n to its class without an error, so that the problem of mean estimation is solved for each class separately, in isolation from the association problem. Therefore, perfectly supervised multiclass learning is no different than a single class problem as far as the learning efficiency is concerned.

Unsupervised or imperfectly supervised learning is essentially different from a single-class problem in that the class associations have to be estimated concurrently with the parameters of the distributions, and the errors in association and parameter estimation contribute to each other. This results in more complicated expressions, which are derived in the following sections.

9.1.4 CRB for General MLANS Concurrent Association and Estimation

In case of multiple modes and classes we number parameters using two indexes, $\mathbf{S}_k = \{S_k^a, a = 1, \ldots, A; k = 1, \ldots, K\}$, a numbers components of observations, and k numbers classes (and modes). The accuracy of estimated parameters $\hat{S}_k^a$ is measured by the covariance matrix, $\mathrm{cov}\left\{\hat{S}_k^a, \hat{S}_{k'}^b\right\}$. Below, for simplicity of notation, we derive explicit expressions for covariances among parameters of a single class, $k' = k$, $\mathrm{cov}\left\{\hat{S}_k^a, \hat{S}_k^b\right\}$. A full covariance including $k \neq k'$ can be computed using the same technique. For the general MLANS model, LL is the logarithm of the AZ-likelihood, AZ-LL, given by Eq. (4-34), and S_k^a include parameters of the model, $\mathbf{M}_k(n, \mathbf{S}_k)$, r_k, and $\mathbf{C}_k$. Using indexes explicitly, $\mathbf{M}_k = \{M_k^i, i = 1, \ldots, I\}$. The derivative $LL_{;a}$ is computed from Eq. (4.3-12) using an identity Eq. (4.2-3),

$$LL_{;a} = \left(\partial/\partial S_k^a\right) \sum_n \ln\left\{\sum_k r_k \ \mathrm{pdf}(\mathbf{x}_n|k)\right\} = \sum_n f(k|n) \ \mathrm{ll}^{;i}(n|k) M_k^{i;a} \tag{9.1-19}$$

here and below, a sum over the repeated indexes i is assumed, and

$$\begin{aligned} f(k|n) &= r_k \ \mathrm{pdf}(\mathbf{x}_n|k) / \sum_{k'} r_{k'} \ \mathrm{pdf}(\mathbf{x}_n|k') \\ \mathrm{ll}^{;i}(n|k) &= \left(\partial/\partial M_k^i\right) \ \ln\left\{r_k \ \mathrm{pdf}(\mathbf{x}_n|k)\right\}, \quad M_k^{i;a} = \partial M_k^i \big/ \partial S_k^a \end{aligned} \tag{9.1-20}$$

In MLANS models, deterministic relationships among various data measurements $\mathbf{x}_n$ are given by models $\mathbf{M}_k$, however, the deviations of data $\mathbf{x}_n$ from the model predictions $\mathbf{M}_k$ are random and uncorrelated. It follows that functions of n, including $f(k|n)$ and $\mathrm{ll}^{;i}(n|k)$, are uncorrelated for different n. Therefore, the covariance in Eq. (9.1-14) is computed as follows:

$$IM^{ab} = \sum_n E\left\{ f^2(k|n)\ \mathrm{ll}^{;i}(n|k)\ \mathrm{ll}^{;j}(n|k) \right\} M_k^{i;a} M_k^{j;b} \tag{9.1-21}$$

Here, sums over the repeated indexes i, j are assumed, and derivatives of means are taken outside the expectation parentheses $E\{\cdot\}$ since they are not random quantities. This expression gives the CRB for the general MLANS model, including MLANS CAT. Let us emphasize that all parameter values in this expression are the true (not estimated) quantities, in particular, $f(k|n)$ are the a posteriori Bayes probabilities.

Let us derive a more specific expression for the CRB, in case of Gaussian class-conditional pdfs, with constant prior rates r_k covariances, $\mathbf{C}_k$, and means $\mathbf{M}_k$ depending on the parameters $\mathbf{S}_k$. Then, the derivatives $\mathrm{ll}^{;i}(n|k)$ can be computed as follows:

$$\begin{aligned} \mathrm{ll}^{;i}(n|k) &= \left(\partial/\partial M_k^i\right)\ \ln\{r_k\ \mathrm{pdf}(\mathbf{x}_n|k)\} = \left(\partial/\partial M_k^i\right)\ \ln\ G[\mathbf{x}_n|\mathbf{M}_k(n,\mathbf{S}_k),\mathbf{C}_k] \\ &= \left(\partial/\partial M_k^i\right)\ \left(-0.5\,\mathbf{D}_{nk}^{\mathrm{T}}\ \mathbf{C}_k^{-1}\ \mathbf{D}_{nk}\right) = D_{nk}^{i'}\left(\mathbf{C}_k^{-1}\right)^{i'i} \end{aligned} \tag{9.1-22}$$

Here, $\mathbf{D}_{nk} = \mathbf{x}_n - \mathbf{M}_k(n,\mathbf{S}_k)$, or using indexes, $D_{nk}^{i'} = x_n^{i'} - M_k^{i'}(n,\mathbf{S}_k)$; a superscript T denotes the transposed vector-column, and a sum over the repeated index i' is assumed. Substituting this into Eq. (9.1-21), we obtain

$$\begin{aligned} IM^{ab} &= \sum_n E\left\{ f^2(k|n)\ D_{nk}^{i'}\left(\mathbf{C}_k^{-1}\right)^{i'i}\ D_{nk}^{j'}\left(\mathbf{C}_k^{-1}\right)^{j'j} \right\} M_k^{i;a} M_k^{j;b} \\ &= \sum_n E\left\{ f^2(k|n)\ D_{nk}^{i'} D_{nk}^{j'} \right\} \left(\mathbf{C}_k^{-1}\right)^{i'i} \left(\mathbf{C}_k^{-1}\right)^{j'j} M_k^{i;a} M_k^{j;b} \end{aligned} \tag{9.1-23}$$

Here, sums over the repeated indexes i, j, i', j' are assumed. This expression can be written in a matrix form:

$$\mathbf{IM} = \sum_n (\partial\mathbf{M}_k/\partial\mathbf{S})^{\mathrm{T}}\ \mathbf{C}_k^{-1}\ E\left\{ f^2(k|n)\ \mathbf{D}_{nk}^{\mathrm{T}}\mathbf{D}_{nk} \right\}\ \mathbf{C}_k^{-1}\ (\partial\mathbf{M}_k/\partial\mathbf{S}) \tag{9.1-24}$$

Equation (9.1-23) or (9.1-24) together with Eq. (9.1-13) give the CRB bound for the case of Gaussian conditional pdfs. More specific, intuitively interpretable expressions are obtained in the following sections.

It is of interest to note that the problem of determining performance limits for concurrent estimation and association (assignment) problems has been a long-standing puzzle.[2] Doubts have been expressed that such bounds could be derived in principle. The confusion has been caused by formulations of the association problem in terms of crisp concepts of Aristotelian logic, while the successful and relatively simple derivation above is due to our reformulation of the association problem in terms of fuzzy adaptive AZ-logic. Although in prediction and tracking, parameters are estimated from multiple time points, numerical evaluation of CRB involves taking only single-pixel expected values and requires integration just over the D-dimensions of single-pixel measurements. The obtained CRB expression can be easily evaluated numerically for any specific functional form of the pdfs and models. In addition to general expressions suitable for numerical evaluation derived above, we will develop in the following sections simplified expressions amenable to intuitive interpretation and suitable for qualitative analysis. In particular, we will analyze in detail the CRB for the statistical

Gaussian mixture model of Chapter 5 (MLANS) and for the dynamic model of concurrent association and tracking of Chapter 7 (MLANS CAT).

9.2 Overlap between Classes

Here we introduce a formalism for the description of overlaps between classes in terms of overlapping and nonoverlapping parts of class populations, means, and covariances. This theory of overlaps is useful for an intuitive interpretation of CRB obtained later.

As discussed previously, the CRB for the means of classes during supervised classification are ellipsoids. We show below that the same is true for the case of unsupervised learning. However, shapes and sizes of the CRB ellipsoids depend on the geometry of the overlap between classes. This result is intuitively clear. Larger overlaps between classes lead to larger association errors and to larger CRB. If a particular class does not overlap with any other class, the CRB for this class is reduced to that of a single-class or a perfectly supervised case. These results are discussed in the following subsections that formally introduce a notion of the distribution overlap, and use it to formulate unsupervised multiclass CRB.

The purpose of this section is to introduce a notion and a quantitative treatment of the overlap between classes. To simplify notations and the discussion, in this section we do not make any distinction between classes and types and we use a single index k to denote classes/types; we also use the same notations $\{r_k, \mathbf{M}_k, \mathbf{C}_k\}$ for the expected and estimated values of these parameters (as long as this does not result in ambiguities). The formalism developed below characterizes overlaps by tensor-like quantities of various orders. A detailed description of the overlap geometry may require the use of high-order quantities. Few lower order quantities of overlapping and nonoverlapping parts of the class populations, means, and covariances introduced here provide approximate description of class overlaps and are useful for the intuitive interpretation of CRB in the next section.

9.2.1 Overlap Matrix

An overlap matrix $O_{kk'}$ is defined as an expected value of the product of fuzzy associations (Bayesian probabilities) for an observation x_n to belong to each of the classes k and k':

$$O_{kk'} \equiv E\{f(k|n) \cdot f(k'|n)\} \tag{9.2-1}$$

Let us remember that $f(k|n)$ is a probability that datum n belongs to class k, so that

$$\sum_k f(k|n) = 1 \tag{9.2-2}$$

For the case of nonoverlapping classes, $O_{kk'}$ is a diagonal matrix:

$$O_{kk'}|_{\mathrm{NO}} = \delta_{kk'} r_k \quad \text{for nonoverlapping case} \tag{9.2-3}$$

This follows from the fact that in a nonoverlapping case $f(k|n)$ equals 0 or 1, therefore $f(k|n)^2 = f(k|n)$, and

$$E\{f(k|n)\} \equiv \int f(k|n)\,\mathrm{pdf}(\mathbf{x}_n)\,d\mathbf{x}_n = \int r_k \cdot \mathrm{pdf}(\mathbf{x}_n|k)\,d\mathbf{x}_n = r_k \tag{9.2-4}$$

An overlap matrix has the property that the sum of row or column elements is equal to the corresponding rate:

$$\sum_{k'} O_{kk'} = r_k \tag{9.2-5}$$

which is a consequence of Eqs. (9.2-2) and (9.2-4). We interpret this equation as an expansion of a rate r_k into its overlapping and nonoverlapping parts. We call $O_{kk'}$ for $k \neq k'$ overlapping parts of r_k. And we call the diagonal element of an overlap matrix, corresponding to the overlap of a class with itself, a nonoverlapping part of a rate:

$$r_k^{\mathrm{NO}} \equiv O_{kk} = r_k - \sum_{k' \neq k} O_{kk'} \tag{9.2-6}$$

A similar expansion for a class population is

$$N_k = \sum_{kk'} \equiv N_{kk'} \equiv N \sum_{k'} O_{kk'}; \quad N_k^{\mathrm{NO}} = N\,O_{kk} \tag{9.2-7}$$

9.2.2 Overlapping Parts of Means

An expansion of a class mean in its overlapping and nonoverlapping parts can be obtained in much the same way using Eqs. (9.2-2) and (5.2-8):

$$\begin{aligned} N_k \cdot \mathbf{M}_k &= E\left\{\sum_n f(k|n)\left[\sum_{k'} f(k'|n)\right] x_n\right\} \\ &= \sum_{k'} E\left\{\sum_n f(k|n)\, f(k'|n)\, x_n\right\} = \sum_{k'} N_{kk'} \cdot \mathbf{M}_{kk'} \end{aligned} \tag{9.2-8}$$

Here, overlapping parts of the mean are defined as

$$\mathbf{M}_{kk'} = E\left\{\sum_n f(k|n)\, f(k'|n)\, x_n\right\} \Big/ N_{kk'} \tag{9.2-9}$$

and a nonoverlapping part of the mean

$$\mathbf{M}_k^{\mathrm{NO}} = E\left\{\sum_n f(k|n)^2\, x_n\right\} \Big/ N_k^{\mathrm{NO}} \tag{9.2-10}$$

9.2.3 Overlapping Parts of Covariance Matrices

Expansion of a covariance matrix in its overlapping parts is a bit more complicated since such an expansion contains in addition to the covariances of the overlapping parts of a class

also a part due to the scatter of the means of the overlapping parts of a class. The deviations of the means are denoted as

$$\Delta\mathbf{M}_{kk'} = \mathbf{M}_{kk'} - \mathbf{M}_k, \quad \Delta\mathbf{M}_k^{\mathrm{NO}} = \mathbf{M}_k^{\mathrm{NO}} - \mathbf{M}_k \tag{9.2-11}$$

and the deviations of the observations $\mathbf{X}_n$ from the overlap means $\mathbf{M}_{kk'}$ are denoted as $\mathbf{D}_{nkk'}$:

$$\mathbf{D}_{nkk'} \equiv \mathbf{X}_n - \mathbf{M}_{kk'} \tag{9.2-12}$$

$$\mathbf{D}_{nk} = \mathbf{D}_{nkk'} + \Delta\mathbf{M}_{kk'} \tag{9.2-13}$$

Proceeding now as in Eqs. (9.2-8), the following expansion for the covariance matrix is obtained:

$$\begin{aligned} N_k \cdot \mathbf{C}_k &= E\left\{\sum_n f(k|n)\,\mathbf{D}_{nk}\,\mathbf{D}_{nk}^{\mathrm{T}}\right\} = E\left\{\sum_n f(k|n)\left[\sum_{k'} f(k'|n)\right]\mathbf{D}_{nk}\,\mathbf{D}_{nk}^{\mathrm{T}}\right\} \\ &= \sum_{k'} E\left\{\sum_n f(k|n)\,f(k'|n)\,\mathbf{D}_{nk}\,\mathbf{D}_{nk}^{\mathrm{T}}\right\} \\ &= \sum_{k'} E\left\{\sum_n f(k|n)\,f(k'|n)\,(\mathbf{D}_{nkk'} + \Delta\mathbf{M}_{kk'})\,(\mathbf{D}_{nkk'} + \Delta\mathbf{M}_{kk'})^{\mathrm{T}}\right\} \end{aligned} \tag{9.2-14}$$

The last line here is obtained by using Eq. (9.2-13). Expanding parentheses in the last line, four items are obtained; of these items, the two cross-items $\sim \mathbf{D}_{nkk'}\Delta\mathbf{M}_{kk'}^{\mathrm{T}}$ and $\sim \Delta\mathbf{M}_{kk'}\mathbf{D}_{nkk'}^{\mathrm{T}}$ are identically zero according to the definitions in Eqs. (9.2-7), (9.2-8), and (9.2-9). Thus,

$$N_k \cdot \mathbf{C}_k = \sum_{k'} N_{kk'} \cdot \mathbf{C}_{kk'} + N_k \cdot \mathbf{C}_k^M \tag{9.2-15}$$

where the overlap-mean covariance is defined as

$$\mathbf{C}_k^M \equiv \sum_{k'} N_{kk'} \cdot \Delta\mathbf{M}_{kk'}\Delta\mathbf{M}_{kk'}^{\mathrm{T}}/N_k \tag{9.2-16}$$

and the covariances of the overlapping and nonoverlapping parts of class k are defined as

$$\mathbf{C}_{kk'} = E\left\{\sum_n f(k|n)\,f(k'|n)\,\mathbf{D}_{nkk'}\,\mathbf{D}_{nkk'}^{\mathrm{T}}\right\}/N_{kk'}; \quad \mathbf{C}_k^{\mathrm{NO}} = \mathbf{C}_{kk} \tag{9.2-17}$$

These definitions are consistent with definitions (9.2-9), (9.2-10), and (9.2-11), so that each of these covariance matrixes can be interpreted as a distribution scatter about the corresponding mean.

A geometric interpretation of the developed formalism is illustrated in Figs. 9.2-1 and 9.2-2, where two-dimensional examples are shown of the distributions of three classes and of their overlapping and nonoverlapping parts. They are plotted here using $2-\sigma$ boundaries of the distributions. The values of parameters of these distributions are summarized in

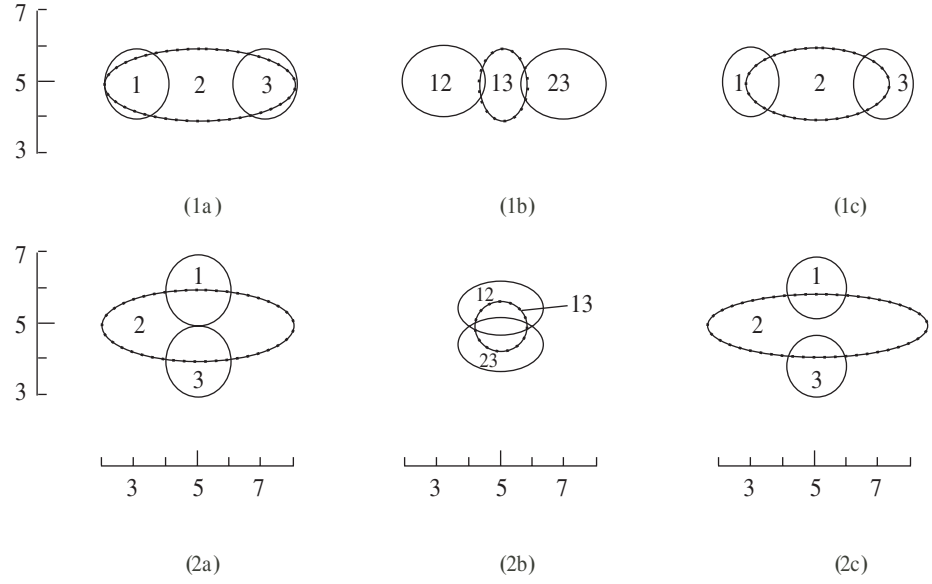


Figure 9.2-1 and **9.2-2** Two sets of distributions for three overlapping classes (**a**), their overlapping parts (**b**), and nonoverlapping parts (**c**) are defined by overlapping and nonoverlapping parts of the means and covariances. They are plotted here using $2-\sigma$ boundaries of the distributions.

the Table 9.2-1. Overlapping parts of the distributions in Figs. 9.2-1b and 9.2-2b and nonoverlapping parts of distributions in Figs. 9.2-1c and 9.2-2c are characterized by their means and covariances. They do not follow exactly the distribution boundaries in Figs. 9.2-1a and 9.2-2a; a more accurate description would require higher order tensors. Note that classes 1 and 3, which are shown separated in Figs. 9.2-1a and 9.2-2a, have small overlapping population. In Figs. 9.2-1b and 9.2-2b, however, the overlapping covariances are relatively large.

Two more quantities useful in the next section are introduced now: second-order statistics:

$$\mathbf{S}_k^{\mathrm{NO}} = E\left\{\sum_n f(k|n)^2\,\mathbf{D}_{nk}\,\mathbf{D}_{nk}^{\mathrm{T}}\right\} \Big/ N_k^{\mathrm{NO}} = \mathbf{C}_k^{\mathrm{NO}} + \Delta\mathbf{M}_k^{\mathrm{NO}}\left(\Delta\mathbf{M}_k^{\mathrm{NO}}\right)^{\mathrm{T}} \tag{9.2-18}$$

TABLE 9.2-1
Parameters of Class Distributions

			Horizontal Component		Vertical Componet		
Example	Class	Prior Probability	Mean	SD	Mean	SD	Correlation
Fig. 9.2-1	1	0.267	3	0.5	5	0.5	0
	2	0.333	5	1.5	5	0.5	0
	3	0.400	7	0.5	5	0.5	0
Fig. 9.2-2	1	0.267	5	0.5	6	0.5	0
	2	0.333	5	1.5	5	0.5	0
	3	0.400	5	0.5	4	0.5	0

and fourth-order statistics:

$$\mathbf{T}_k^{\mathrm{NO}} = E\left\{\sum_n f(k|n)^2\, \mathbf{D}_{nk} * \mathbf{D}_{nk} * \mathbf{D}_{nk} * \mathbf{D}_{nk}\right\} / N_k^{\mathrm{NO}} \tag{9.2-19}$$

An expansion for $\mathbf{T}_k^{\mathrm{NO}}$ similar to the one obtained for $\mathbf{S}_k^{\mathrm{NO}}$ can be developed. However, an intuitive appeal driving the development in this section is decreasing when dealing with a fourth-order tensor. The definition (9.2-19) can be used for numerical calculations of the CRB for covariance estimation obtained in the next section.

9.3 CRB FOR MLANS

Lower bounds for the variances of the estimated parameters of a mixture are evaluated in this section according to the CR theorem, and expressed in terms of overlaps between classes. These expressions are intuitively appealing and also can be easily evaluated numerically for any given probability distribution. It is necessary to distinguish in this section parameters of distributions from their estimated values, which are denoted by a "hat."

9.3.1 CRB for Prior Rates

The derivation of the CRB for prior rates is complicated by interdependence between rates due to the normalization constraint $r_k = 1$. A derivative $\partial/\partial r_k$ can be evaluated along different paths in $\{r_k\}$ space satisfying this constraint. To obtain a correct definition of this derivative for our purpose, it should be remembered that the CRB is given by the inverse variance of the log-likelihood, derivative. The CRB is the tightest bound (that is the uppermost lower bound) among various bounds associated with different definitions of $\partial/\partial r_k$ derivative. Thus, the CRB corresponds to the minimal variation in $\partial LL/\partial r_k$ and the derivative should be defined as to minimize $\mathrm{var}(\partial LL/\partial r_k)$.

A general definition of this derivative can be obtained by varying all the $r_{k'}, k' = 1, \ldots, K$ independently under the normalization constraint:

$$r_{k'} \to r_{k'} + \alpha_{k'} \Delta r_k, \quad k' = 1, \ldots, K, \quad \sum_{k'} \alpha_{k'} = 0, \quad \alpha_k = 1 \tag{9.3-1}$$

The last constraint here is added for convenience and does not reduce the generality of this definition. The corresponding variation in LL is

$$\Delta LL = \partial LL/\partial r_k \cdot \Delta r_k = \sum_n \sum_{k'} f(k'|n)\, [\alpha_{k'}/r_{k'}]\, \Delta r_k \tag{9.3-2}$$

which yields the derivative

$$\partial LL/\partial r_k \equiv \Delta LL/\Delta r_k = \sum_n \sum_{k'} f(k'|n)\, [\alpha_{k'}/r_{k'}] \tag{9.3-3}$$

The variance of this expression can be evaluated as follows [remember that the expected value of (9.3-3) is zero and for different n, $f(k|n)$ are independent]:

$$\begin{aligned}\mathrm{var}\{\partial LL/\partial r_k\} &= \sum_n \mathrm{var}\left\{\sum_{k'} f(k'|n)\,[\alpha_{k'}/r_{k'}]\right\} \\ &= N\cdot E\left\{\sum_{k';k''} f(k'|n)f(k''|n)\;\;[\alpha_{k'}\alpha_{k''}/r_{k'}r_{k''}]\right\} \\ &= N\sum_{k';k''} O_{k'k''}\;\;[\alpha_{k'}\alpha_{k''}/r_{k'}r_{k''}]\end{aligned} \tag{9.3-4}$$

An appropriate definition of the derivative $\partial/\partial r_k$ leading to the CRB is obtained now by minimizing this expression over the set of $\{\alpha_k\}$ under the constraints (9.3-1). A Lagrangian multiplier method results in the following set of equations:

$$\begin{aligned}&2\sum_{k';k''} O_{k'k''}\,(\alpha_{k'}/r_{k'}r_{k''}) + \lambda_0 + \lambda_k\delta_{kk''} = 0, \quad k'' = 1,\ldots,K; \\ &\sum_{k'} \alpha_{k'} = 0; \quad \alpha_k = 1\end{aligned} \tag{9.3-5}$$

This set of $K+2$ linear equations for $K+2$ variables $\{\alpha_k, \lambda_0, \lambda_k\}$ can be solved numerically for any given model. Numerical evaluation of an overlap matrix $O_{k'k''}$ involves just one D-dimensional integration. Resulting values of α_k should be substituted into Eq. (9.3-4) yielding the CRB for the MLANS r_k estimator:

$$\mathrm{var}\left\{\hat{r}_k\right\} \geq \left[N\sum_{k';k''} O_{k'k''}\,(\alpha_{k'}\alpha_{k''}/r_{k'}r_{k''})\right]^{-1} \tag{9.3-6}$$

A simple solution of (9.3-6) can be obtained in the case when class k overlaps with a single class k', much stronger than with any other class. It is sufficient then to consider just these two classes, $\alpha_{k'} = -\alpha_k = -1$, leading to the CRB:

$$\begin{aligned}\mathrm{var}\left\{\hat{r}_k\right\} &\geq N^{-1}\,E\left\{\left[f(k|n)/r_k - f(k'|n)/r_{k'}\right]^2\right\}^{-1} \\ &= N^{-1}\,\left[O_{kk}/r_k^2 + O_{k'k'}/r_{k'}^2 - 2O_{kk'}/r_k r_{k'}\right]^{-1}\end{aligned} \tag{9.3-7}$$

The maximum variance here corresponds to two identical classes k and k' (maximum overlap); if the overlap is insignificant this expression is reduced to

$$\mathrm{var}\left\{\hat{r}_k\right\} \geq N^{-1}\,[1/r_k + 1/r_{k'}]^{-1} = N^{-1}\,[(r_k + r_{k'})/(r_k r_{k'})]^{-1} \tag{9.3-8}$$

For the two-class case, $r_k + r_{k'} = 1$, and for a nonoverlapping class:

$$\mathrm{var}\left\{\hat{r}_k\right\} \geq N^{-1}\,r_k\,(1 - r_k) \tag{9.3-9}$$

The same expression can be obtained in a case of no overlap between all classes. In this case the overlap matrix is given by Eq. (9.2-3), and it can be verified by the substitution that the following is a solution of Eqs. (9.3-5):

$$\alpha_{k'} = (\delta_{kk'} - r_{k'})/(1 - r_k) \tag{9.3-10}$$

This leads to the following CRB for the no-overlap case:

$$\text{var}\left\{\hat{r}_k\right\} \geq N^{-1}\, r_k(1 - r_k) \tag{9.3-11}$$

It is interesting to note, that in nonoverlapping cases the CRB does not go to 0, as in a single class case; the reason is that in a multiclass case there remains a randomness of r_k.

9.3.2 CRB for Means

The CRB for means is obtained directly from (9.1-23). In this case, a set of S^a is identical to a set of M^i, therefore, $M_k^{i;a} = \delta^{i;a}$, and $IM^{ab} = IM^{ij}$,

$$\begin{aligned} IM^{ij} &= \sum_n E\left\{f(k|n)^2\, D_{nk}^{i'}\, D_{nk}^{j'}\right\} \left(\mathbf{C}_k^{-1}\right)^{i'i} \left(\mathbf{C}_k^{-1}\right)^{j'j}; \quad \text{or} \\ \mathbf{IM} &= N_k^{\text{NO}}\, \mathbf{C}_k^{-1}\, \mathbf{S}_k^{\text{NO}}\, \mathbf{C}_k^{-1} \end{aligned} \tag{9.3-12}$$

where $\mathbf{S}_k^{\text{NO}}$ has been defined in Eq. (9.2-18). This yields the following CRB for the mean components of the kth class,

$$\text{cov}\left\{\hat{\mathbf{M}}_k, \hat{\mathbf{M}}_k\right\} \geq \left(N_k^{\text{NO}}\right)^{-1}\, \mathbf{C}_k\, \left(\mathbf{S}_k^{\text{NO}}\right)^{-1}\, \mathbf{C}_k \tag{9.3-13}$$

This result is intuitively appealing: a large CRB and large errors correspond to a large overlap (small nonoverlapping part of a class). If class k does not overlap with any other class, this expression is reduced to the usual covariance of the mean estimator.

Figures 9.3-1 and 9.3-2 show CRB for means for the examples considered above in Figs. 9.2-1 and 9.2-2, respectively. Here the CRB are multiplied by 2 in order to be comparable with $2 - \sigma$ distributions in Figs. 9.2-1 and 9.2-2. In addition, the CRB in Figs. 9.3-1a and 9.3-2a are multiplied by the number of observations in order to remove this dependence. Their shape is different from that in Figs. 9.2-1a and 9.2-2a due to the nonoverlapping parts of covariances shown in Figs. 9.2-1c and 9.2-2c, and their size is determined by nonoverlapping parts of the class populations. In Figs. 9.2-1b and 9.2-2b the same CRB are shown multiplied by the corresponding number of observations in each class; this removes the effects of rates and these ellipses, when compared with the distributions in Figs. 9.2-1a and 9.2-2a, show the CRB increase due to classification errors.

9.3.3 CRB for Covariances

The CRB for an inverse covariance $\mathbf{C}_k^{-1}$ is obtained below, because it has a somewhat simpler form than the one for $\mathbf{C}_k$. The derivatives are calculated as

$$\begin{aligned} \partial LL/\partial \mathbf{C}_{kab}^{-1} = \sum_n f(k|n)\, (&-\mathbf{D}_{nka}\mathbf{D}_{nkb} + 1/2\, \delta_{ab}\mathbf{D}_{nka}\mathbf{D}_{nka} \\ &+ \mathbf{C}_{kab} - 1/2\, \delta_{ab}\mathbf{C}_{kaa}) \end{aligned} \tag{9.3-14}$$

Using definitions (9.2-18) and (9.2-19), the fourth-order tensor of covariances of these derivatives is evaluated as follows:

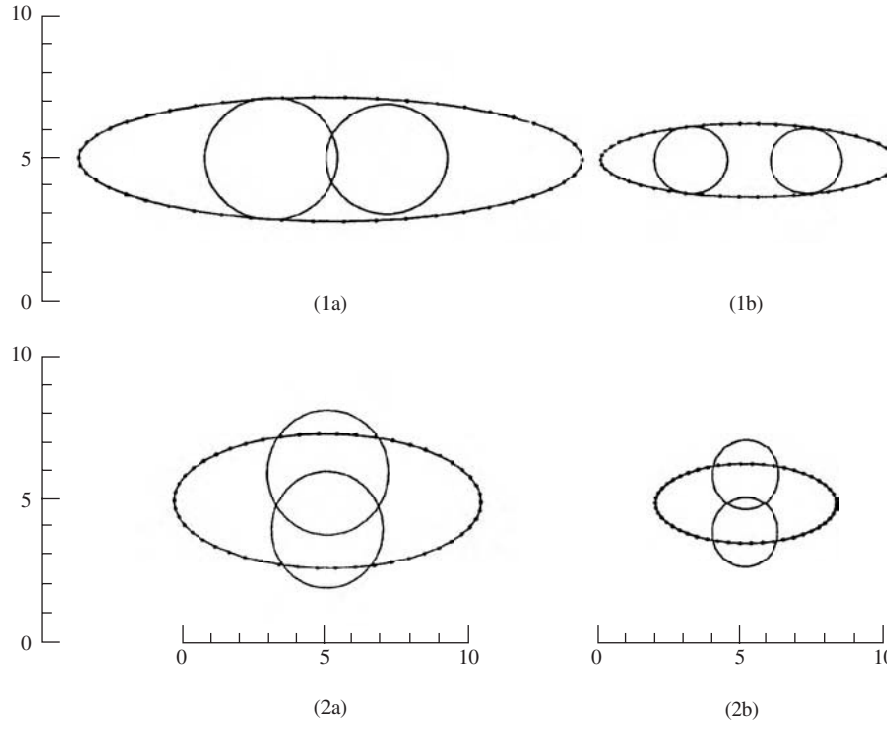


Figure 9.3-1 and **9.3-2** Cramer–Rao bounds for means of overlapping classes shown in Figs. 9.2-1 and 9.2-2.

$$
\begin{aligned}
&\operatorname{cov}\left\{\partial LL/\partial \mathbf{C}^{-1}_{kii'}\ \partial LL/\partial \mathbf{C}^{-1}_{kjj'}\right\} \\
&= N_k^{\mathrm{NO}}\big(\mathbf{T}^{\mathrm{NO}}_{kii'jj'} - 1/2\,\delta_{ii'}\mathbf{T}^{\mathrm{NO}}_{kiijj'} - 1/2\,\delta_{jj'}\mathbf{T}^{\mathrm{NO}}_{kii'jj} + 1/4\,\delta_{ii'}\delta_{jj'}\mathbf{T}^{\mathrm{NO}}_{kiijj} \\
&\quad - \mathbf{S}^{\mathrm{NO}}_{kii'}\,\mathbf{C}_{kjj'} - \mathbf{S}^{\mathrm{NO}}_{kjj'}\,\mathbf{C}_{kii'} + 1/2\,\delta_{jj'}\mathbf{S}^{\mathrm{NO}}_{kii'}\,\mathbf{C}_{kjj} + 1/2\,\delta_{ii'}\mathbf{S}^{\mathrm{NO}}_{kjj'}\,\mathbf{C}_{kii} \\
&\quad + \mathbf{C}_{kii'}\mathbf{C}_{kjj'} - 1/2\,\delta_{ii'}\mathbf{C}_{kii}\mathbf{C}_{kjj'} - 1/2\,\delta_{jj'}\mathbf{C}_{kii'}\mathbf{C}_{kjj} + 1/4\,\delta_{ii'}d_{jj'}\mathbf{C}_{kii}\mathbf{C}_{kjj}\big)
\end{aligned}
\tag{9.3-15}
$$

A summation over repeated indices is *not* assumed here. The inverse of this expression yields the CRB for the estimated components of the covariance matrix of class k. This expression is too complicated for an intuitive appeal; still it is suitable for a numerical evaluation.

9.3.4 MLANS Performance vs. CRB: Example 3 Continuation

In Section 5.2.3 we compared the classification performance of the MLANS and ISODATA algorithm using a standard data set. Here, continuing this example, we compare the MLANS performance vs. CRB. In Section 5.2.3, we used the Bhattacharyya distance to characterize

the difference between the true and estimated distributions (the Bhattacharyya distance, in turn, is related to classification errors). We discussed that Bhattacharyya goes to zero when a number of data samples available for the estimation goes to infinity. For the finite number of samples, the estimation error is finite and the Bhattacharyya is nonzero, positive. The minimal expected deviation of Bhattacharyya from zero, or its minimal bias, can be computed from the CRB for the parameters of the distribution. Fukunaga and Hayes (1988) have computed this minimal bias for the case of a single Gaussian distribution, or equivalently, for the case of perfect supervision.

This minimal bias is shown in Fig. 9.3-3 using the dashed line; the MLANS performance is shown using solid lines. The MLANS performance in this case is as good as the Fukunaga–Hayes bound. Two points should be further clarified. First, the MLANS performance is sometimes better than the bound (below the dashed line). This does not contradict the nature of the bound: the dashed line is the bound for the best *average* performance of any algorithm or neural network. Second, MLANS reaches this bound without any knowledge of class assignments, although the bound is derived given the perfect knowledge of class assignments. Although this fact is remarkable, it can be explained as follows. From the previous section, Eqs. (9.3-6), (9.3-13), and (9.3-15), it follows that the CRB for the parameters of mixture components are affected by the overlap between components; in a case of no overlap, $f(k|n) = 1$ or 0, $f(k|n)^2 = f(k|n) = 1$ or 0, and the CRB are no different from the perfectly supervised case. Since in the considered example the overlap is small (the Bayes error is $= 1.9\%$), the CRB are close to the perfectly supervised case.

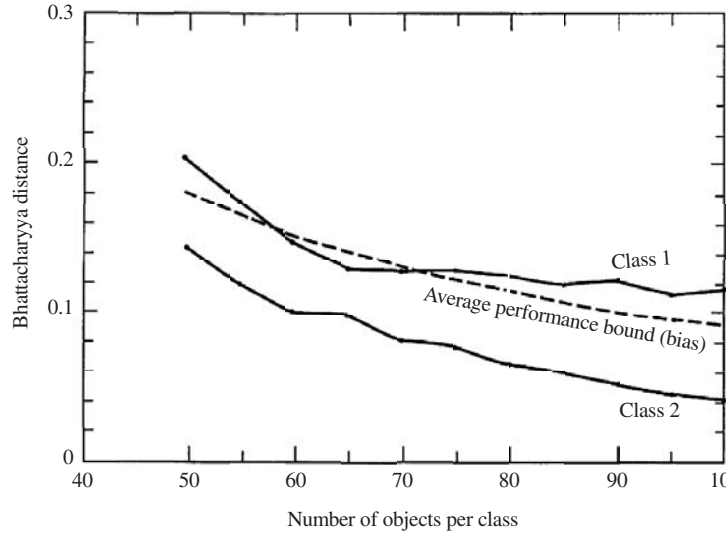


Figure 9.3-3 Comparison of the MLANS performance vs. the information-theoretic bound. Example 3 continuation, same as in Fig. 5.2-8; a blow up of the region $N \geq 50$. The dashed line shows the theoretical bound for an average performance of any algorithm calculated for the perfectly supervised case (same for each class).

One of the consequences of the CR theorem, mentioned in Chapter 4, is that the ML estimation asymptotically achieves the CRB. The above example demonstrates that, in agreement with this general statement, the actual learning efficiency of the MLANS in this case has achieved the theoretical limit of the CRB.

9.4 CRB for Concurrent Association and Tracking (CAT)

Performance limits for tracking in clutter, previously an unsolved problem, were derived only recently (see Note 2). In this section we obtain the CRB for tracking as a consequence of the general CRB for model-based association and estimation. Then we derive simplified rule-of-thumb expressions for qualitative analysis and discuss their intuitive interpretation.

9.4.1 CRB for Linear Tracks

The CRB for the CAT problem is given by Eqs. (9.1-13), (9.1-21), and (9.1-23), when track models are substituted in these expressions. The track models were discussed in Chapter 7. Now consider the CR bound for a case when (1) the distributions of classification features are independent of the position and motion of objects, and (2) the object of interest is on a linear track and there is no Doppler measurements. So the measurement vector for each pixel or sample consists of coordinates and classification features, $\mathbf{X}_n = (\mathbf{Y}_n, \mathbf{R}_n)$. And the conditional pdf of the nth measurement, under the hypothesis that it came from an object on track k, factors into a product of the positional and classification parts:

$$\text{pdf}(\mathbf{Y}_n, \mathbf{R}_n | H_k) = \text{pdf}(\mathbf{R}_n | H_k,\ \text{tracking}) \cdot \text{pdf}(\mathbf{Y}_n | H_k,\ \text{classification}) \qquad (9.4\text{-}1)$$

The expected value of the position of object k, at time t_n, $\mathbf{M}_k(t_n)$ is given by the linear track model, $\mathbf{M}_k = \mathbf{R}_k + \mathbf{V}_k t_n$. Parameters of this model are $\mathbf{S}_k = (\mathbf{R}_k, \mathbf{V}_k)$. The deviations of pixel positions from the model predictions are $\mathbf{D}_{nk} = \mathbf{R}_n - \mathbf{M}_k(t_n)$. And pdf of these deviations are modeled using Gaussian distributions

$$\text{pdf}(\mathbf{R}_n | H_k,\ \text{tracking}) = G\,[\mathbf{R}_n | \mathbf{M}_k(t_n), \mathbf{C}_k] \qquad (9.4\text{-}2)$$

The model derivatives are computed in Problem 9.4–1 (which also contains the details of the following derivations). Substituting these derivatives into Eqs. (9.1-19) and (9.1-21) we obtain

$$\begin{aligned} \partial LL/\partial \mathbf{S}_k &= \sum_n f^2(k|n)\ \mathbf{D}_{nk}\ \mathbf{C}_k^{-1}\ (1, t_n)^{\mathrm{T}} \\ \mathbf{IM} &= \sum_n \begin{bmatrix} 1 & t_n \\ t_n & t_n^2 \end{bmatrix} \mathbf{C}_k^{-1}\ E\left\{ f^2(k|n)\ \mathbf{D}_{nk}^T \mathbf{D}_{nk} \right\} \mathbf{C}_k^{-1} \end{aligned} \qquad (9.4\text{-}3)$$

The inverse of this expression gives the CRB for concurrent association and tracking of multiple targets on linear trajectories in clutter. It is a 4×4 matrix corresponding to the four track model parameters: positions and velocities for each of the two coordinates. This 4×4 matrix is written here as a cross product of two 2×2 matrixes. The first 2×2 matrix under

the sum in the square brackets corresponds to position and velocity parameters. The 1 in the upper diagonal element in the matrix gives the error bound for the target position estimation, $\mathbf{R}_k$. The lower diagonal element t_n^2 gives the error bound for the target velocity estimation, $\mathbf{V}_k$. The off-diagonal element in this matrix gives the covariances between errors in position and velocity. The rest of this expression is a 2×2 matrix of coordinates (range and angle). Expression (9.4-3) can be numerically evaluated for any specific target motion, clutter, and classification feature distributions to obtain the fundamental limitations on the accuracy of tracking. This expression can also be interpreted in terms of class overlap similar to Eq. (9.3-13).

When Doppler measurements of the radial velocity are available, the data for each sample n contain the Doppler velocity V_n in addition to the radial position R_n and angle position, $\mathbf{R}_{\Theta n} : (\mathbf{R}_{\Theta n}, R_n, V_n)$. The CRB for the angular components are not affected, but the CRB for the radial velocity V_k is different from Eq. (9.4-3) and is obtained from the general CR bound Eq. (9.1-21) as follows. The radial component model is given by $\mathbf{M}_k = (R_k + V_k t_n, V_k)$. The radial (position and velocity) components of the model derivatives $\mathbf{M}_{k;v}$ and deviations of the data from model $\mathbf{D}_{nk}$ are

$$\begin{aligned} &\mathbf{M}_{k;v} = (t_n, 1) \\ &\mathbf{D}_{nk} = (DR_{nk}, DV_{nk}),\ DR_{nk} = R_n - R_k - V_k t_n, \quad DV_{nk} = V_n - V_k \end{aligned} \tag{9.4-4}$$

The CR bound is obtained by substituting this into Eq. (9.1-21). If the measurement errors in the radial position and velocity are independent,

$$\mathbf{C}_k^{-1} = \begin{bmatrix} \sigma_R^{-2} & 0 \\ 0 & \sigma_V^{-2} \end{bmatrix} \tag{9.4-5}$$

the CR bound for the velocity is

$$\sigma_{V,CR}^{-2} = \sum_n E\left\{ f^2(k|n) \left[t_n\ DR_{nk}\ \sigma_R^{-2} + DV_{nk}\ \sigma_V^{-2} \right]^2 \right\} \tag{9.4-6}$$

Usually, Doppler measurements are much more accurate than range measurements, so

$$\begin{aligned} &t_n\ DR_{nk}\ \sigma_R^{-2} \ll DV_{nk}\ \sigma_V^{-2} \\ &\sigma_{V,CR}^{-2} \approx \sum_n E\left\{ f^2(k|n) DV_{nk}^2\ \sigma_V^{-4} \right\} = \sigma_V^{-4} \sum_n E\left\{ f^2(k|n)\ DV_{nk}^2 \right\} \end{aligned} \tag{9.4-7}$$

This expression can be interpreted in terms of class overlap similar to Eq. (9.3-13).

9.4.2 Rule-of-Thumb CRB for CAT

Now, we obtain and analyze a simplified expression from Eq. (9.4-3), that will give a "rule of thumb" for the accuracy of tracking in clutter. Summary results are discussed below; a complete derivation is presented in the Appendix (Section 9.6). For this simplified analysis, we consider the CR bound for the position parameter (range or angle) of the object-track model, σ_{CR}^2. In terms of the matrix elements of Eq. (9.4-3), this bound corresponds to the upper diagonal element of the first matrix in Eq. (9.4-3), which equals 1, and the

upper diagonal element of matrix $\mathbf{C}_k^{-1}$, denoted σ^{-2}, which is the standard deviation of the sensor position measurement determined by the sensor accuracy, pointing precision, object apparent size (e.g., length of the return signal), signal propagation environment, etc.

The "rule of thumb" simplified expression is obtained in terms of several parameters that include the number of observations over time used to initiate or estimate track N_t (this is the number of scans, or frames, or pings), the number of pixels per track accuracy N_p ($N_p \sim 2\sigma\tau^{-1}$, where τ is a pixel size), signal-to-clutter ratio S/C, and "detectability" (or "classifiability") of a single pixel: an average likelihood ratio of classification features of a single object pixel, LRCF_O (exact definitions are given in the Appendix). The CR bound intuitive interpretation is discussed below in terms of the S/C ratio, even though neither the CR bound nor rules of thumb derived below depend on any kind of signal thresholding procedure.

In a case of no clutter (perfect object-pixel detection/classification), tracking accuracy improves with the number of frames N_t used for track estimation,

$$\sigma_{\text{CR}} \sim \sigma N_t^{-1/2}, \quad \text{no clutter} \tag{9.4-8}$$

This is a familiar $\sqrt{N_t}$ law for the accuracy of repeated measurements, as should be expected.

In a case of very strong clutter, (signal-to-clutter ratio, S/C $\to$ 0),

$$\sigma_{\text{CR}} \sim \sigma N_t^{-1/2} N_p^{1/2}, \quad \text{for very strong clutter} \tag{9.4-9}$$

This expression can be used for a qualitative analysis of tracking requirements. Successful tracking requires that the accuracy of the track improves (does not degrade) comparatively to the accuracy of a single position measurement. Thus for a successful tracking, $\sigma_{\text{CR}}^2 < \sigma^2$, or

$$N_t^{-1}\, N_p < 1 \quad \text{or} \quad N_t > N_p, \quad \text{for very strong clutter} \tag{9.4-10}$$

For a typical example characterized by $N_p \sim 100$, to satisfy Eq. (9.4-10), the number of frames used to initiate tracks should be $N_t > 100$ for very strong clutter. This is often prohibitive. Still, using high frame-rate sensors, satisfying this requirement, it is possible to initiate tracks for S/C < 1 by using CAT procedure with more than 100 frames. As expected, tracking in very strong clutter can be accomplished either by initiating tracks on a very large number of frames, or by concurrent utilization of classification features. Utilization of classification features is equivalent to S/C reduction; this is explicitly shown below.

Let us derive the CRB rule of thumb in a case of strong clutter, but not very strong S/C $\lesssim$1 (or less if classification features are used). To obtain expressions amenable to intuitive interpretation, we separate classification features into a signal-to-clutter ratio, S/C, and other features as might be available. The rule of thumb below is obtained assuming that a signal coming from an object has a constant amplitude S and, in addition, there is an additive random clutter; clutter is characterized by its standard deviation, C. The CRB rule of thumb is obtained in terms of the distribution parameters of average and peak clutter. The reason for this is that in a pixel (or sample) containing signal and clutter, the signal + clutter distribution is determined by statistical properties of an average clutter pixel, but the distribution of clutter alone is determined by statistical properties of a peak clutter pixel within the sensor accuracy window N_p. Therefore, the rule-of-thumb expression below is derived in terms of the means and standard deviations of an average clutter sample, m_c and σ_c, and of a peak

clutter pixel within a sensor accuracy window N_p, m_p, and σ_p. In addition to S/C, other available classification features can be used for association and tracking; these features are accounted for by using $\mathrm{LRCF_O}$: an average likelihood ratio of classification features of a single object pixel, excluding an S/C feature (exact definition is given in the Appendix):

$$\begin{aligned} &\sigma_{\mathrm{CR}} \sim \sigma N_t^{-1/2} \left\{1 + N_p \exp\left[-(S+B)/\sigma_p\right] \cdot \mathrm{LRCF_O}^{-1}\right\} \\ &B = m_c - m_p + \sigma_p - \sigma_c \end{aligned} \tag{9.4-11}$$

In the above equation, the likelihood ratio $\mathrm{LRCF_O}$ is multiplied by $\exp(S/\sigma_p)$. This illustrates an intuitively obvious point that classification features can be used to "improve" the S/C ratio. In fact, an "effective S/C" can be defined as

$$\text{"effective S/C"} = \mathrm{S/C} + (\sigma_p/C) \ln\ \mathrm{LRCF_O} \tag{9.4-12}$$

(this expression is valid for any definition of a clutter measure for C, not just the standard deviation). Equation (9.4-12) is applicable in case of weak clutter as well. By comparing Eqs. (9.4-12) and (9.4-8), we obtain a definition of what is a weak clutter [in the context of Eq. (9.4-8)]:

$$\exp\left[(S+B)/\sigma_p\right] \cdot \mathrm{LRCF_O} > N_p \tag{9.4-13}$$

Let us require, as before, that for a successful tracking, $\sigma_{\mathrm{CR}} < \sigma$. Then, the requirement on S/C for tracking follows from Eq. (9.4-12):

$$\begin{aligned} &\exp\left[(S+B)/\sigma_p\right] \cdot \mathrm{LRCF_O} \gtrsim N_t^{-1/2} N_p \quad \text{or} \\ &S + \sigma_p \ln\ \mathrm{LRCF_O} \gtrsim -B + \sigma_p\ \ln\left(N_p/N_t^{1/2}\right) \end{aligned} \tag{9.4-14}$$

This can be rewritten so that any favorite definitions of clutter measure C can be used:

$$\mathrm{S/C} + (\sigma_p/C)\ \ln\ \mathrm{LRCF_O} \gtrsim -B/C + (\sigma_p/C)\ \ln\left(N_p/N_t^{1/2}\right) \tag{9.4-15}$$

This gives a rule of thumb for tracking in clutter.

Let us explore some typical numerical values of the parameters in Eqs. (9.4-12) and (9.4-15). Parameters of distributions of clutter peak can be derived, assuming particular pdf for average clutter. For exponential and uniform distributions results are summarized in Table 9.4-1. Table 9.4-1 also contains parameters estimated from some typical real sensory data. These are shown for illustrative purpose only; of course, properties of real data vary from case to case and from sensor to sensor.

For the example of our sensor data, $N_p = 100$, without using any other classification features, the requirement for track initiation with four frames, $N_t = 4$,

$$\mathrm{S}/\sigma_c \gtrsim 3.0 \quad (\sim 5 \text{ db, or less if classification features are used}) \tag{9.4-16}$$

This is a moderate S/C requirement. If other classification features are used, $\mathrm{LRCF_O} > 1$ and S/C requirements are even lower.

Expressions (9.4-10), (9.4-11), (9.4-12), and (9.4-15) give surprisingly simple rules of thumb for the possibility of tracking in clutter in terms of the number of frames, S/C ratio,

TABLE 9.4-1
Parameters of Distributions

Distribution	$\mathbf{m_c}/\sigma_\mathbf{c}$	$\sigma_\mathbf{c}$	$\mathbf{m_p}/\sigma_\mathbf{c}$	$\sigma_\mathbf{p}/\sigma_\mathbf{c}$	$\mathbf{B}/\sigma_\mathbf{c}$
$\exp(-x/\sigma_c)/\sigma_c$			$N_p = 10, 50, 100, 200, 300, 5000$		$N_p = 100$
	1	σ_c	$m_p = 3, 4.6, 5.2, 5.9, 6.3, 9$	1.3	-3.9
$\mathrm{U}\left[0, 2\sqrt{3}\cdot\sigma_c\right]$	$\sqrt{3}$	σ_c	$2\sqrt{3}\ \left(1 - N_p^{-1}\right)$	N_p^{-1}	$N_p = 100$
			≈ 3.4		-2.7
Sensor data	0.57	1.75	1.71	0.29	-1.85

sensor accuracy, and single-frame classification likelihood. Using this result, the possibility of tracking in clutter can be evaluated by studying statistical distributions of single pixel measurements of single-frame features.

9.5 Summary: CRB for Intellect and Evolution?

Let us summarize this chapter. The CRB is a fundamental mathematical limit for learning of any adaptive or learning algorithm or neural network. This limit is inherent to the limited amount of information present in data and a priori models. We derived the CRB for the problem of concurrent association and estimation. In particular, two problems were considered in detail: estimation of the normal mixture pdf (clustering, or concurrent association and estimation) and concurrent association, detection, and tracking for targets in clutter. The Cramer–Rao bounds were formulated and closed-form expressions were derived suitable for numeric evaluation. We obtained intuitively interpretable expressions for CRB using the developed formalism of class overlaps and rule-of-thumb approximate CRB expressions were obtained that provide for a qualitative, intuitively appealing characterization of the possibility of tracking in clutter. The developed technique can easily be extended to obtain more general results such as CRB for partial or imperfect supervision.

We derived intuitively interpretable expressions and rules of thumb. The CRB for the means of the mixture components has a simple intuitive interpretation. The shape of the bound is ellipsoidal, the same as in the case of a single Gaussian distribution. The size of this ellipse, as expected, is increased by overlaps between mixture components. In case of no overlap, the bound is exactly same as for a single Gaussian distribution. Rules of thumb for tracking in clutter turned out to be surprisingly simple expressions in terms of the number of frames (or scans), S/C ratio, and single-pixel classification likelihood. The derived rules of thumb indicate that it should be possible to track objects in a relatively low signal-to-clutter environment ($\sim$ 5 db for unresolved objects in random clutter), and even this moderate requirement can be significantly relaxed if classification features are used while performing concurrent association, classification, and tracking, or when positional errors of measurements are small. Using these results, the possibility of tracking in clutter can be evaluated by studying statistical distributions of single-pixel signals or single-frame features.

For a long time CRB has been available for relatively simple problems in signal processing and other areas of statistical estimation. It has been used for levying sensor requirements, for the diagnostics of algorithms that potentially can be improved, and for determination of

how many data are needed to obtain an estimation of parameters in particular problems. In the same way the CRB described here can be used for the diagnostics of learning systems performing concurrent classification, association, and tracking. We illustrated use of the CRB for performance evaluation. Other uses of the CRB in neural network development may include the following. The CRB depends on various problem parameters, such as the number of classes and types and the dimension of the classification space. Therefore, it can be used for estimating the best possible scaling properties of neural networks and intelligent systems, when projecting the requirements for large-scale systems. It will be important to understand what can be deduced from the CRB for complicated very high-dimensional problems such as image understanding, where the number of training samples is smaller than the number of pixels, and therefore a meaningful estimation is possible only with the help of additional, model-based information. Which additional information is important, and what should be the properties of efficient a priori models?

Are CRB applicable to the intellect in general? Given an important role of internal models within intelligent systems, could CRB be used for establishing limits on speed of learning and evolution of any intelligent system, such as humankind? There seems to be one main difficulty for such a project: a CR's assumption that the structure and functional shape of the internal model is available a priori so that the learning and evolution can be described as a parameter estimation. To overcome this difficulty, let us consider the genetic information determining the internal model as a part of the model. And consider the genetic evolution as a part of the model estimation. Thus, the structure and functional shape of the internal model are parameterized (in terms of the genetic information), and the entire evolution can be viewed as an adaptation–estimation of the internal model. Applying the CR theory to this evolution-learning process involves another difficulty: the genetic information is discrete, whereas the CR theory requires continuous parametric model and evaluation of the derivatives. Could this be resolved by considering the internal model of the entire human genome and its parameters to be average characteristics of the population? In Chapter 10 we make a step toward mathematical formulation of the notion of evolutionary gradient acting on the genome. So the Cramer–Rao bound for the entire process of evolution seems to be an interesting problem for future research.

9.6 Appendix: CRB Rule of Thumb for Tracking

Details of the derivation of the rule of thumb for tracking in clutter are discussed here. The rule of thumb is derived from the exact CR bounds, Eq. (9.4-3), for a case of a single moving object in random clutter. The expected value in Eq. (9.4-3) can be approximately evaluated as

$$E\left\{f^2(k|n)\ \mathbf{D}_{nk}^T\mathbf{D}_{nk}\right\} \sim E\left\{f^2(k|n)\right\}\ E\left\{\mathbf{D}_{nk}^{\mathrm{T}}\mathbf{D}_{nk}\right\} \tag{A9-1}$$

The probabilities $f(k|n)$ and deviations $\mathbf{D}$ are not statistically independent and this approximation is valid only as an order-of-magnitude estimate. The second term here, by definition, is the covariance matrix,

$$E\left\{\mathbf{D}_{nk}^{\mathrm{T}}\mathbf{D}_{nk}\right\} = \mathbf{C}_k \tag{A9-2}$$

Thus, we obtain

$$\mathbf{IM} = \sum_n \begin{bmatrix} 1 & t_n \\ t_n & t_n^2 \end{bmatrix} \mathbf{C}_k^{-1} \; E\left\{f^2(k|n)\right\} \tag{A9-3}$$

Consider the CR bound for the position parameter of the object-track model, σ_{CR}^2. This bound is given by the upper diagonal element of the matrix in brackets, which equals 1, and the upper diagonal element of matrix $\mathbf{C}_k^{-1}$, denoted σ^{-2}, which is the standard deviation of the sensor position measurement,

$$\sigma_{\text{CR}}^{-2} \sim \sigma^{-2} \sum_n \; E\left\{f^2(k|n)\right\} \tag{A9-4}$$

The expected value here can be approximated as follows. For computation of the probabilities, we simplify expressions for classification and tracking pdfs for two classes: object and clutter. We substitute the Gaussian distribution of object tracking errors with a uniform distribution centered at the expected object position and having a width determined by the track model error, σ. So, the object-track pdf is given by

$$\begin{aligned} &\text{pdf}(\mathbf{R}_n|H = \text{object, tracking}) = 1/(2\sigma), \quad \text{if} |\mathbf{M}_k(t_n) - \mathbf{R}_n| < \sigma; \\ &\text{otherwise} = 0 \end{aligned} \tag{A9-5}$$

We consider random clutter, uniformly distributed in range, so clutter-track pdf is given by

$$\text{pdf}(\mathbf{R}_n|H = \text{clutter, tracking}) = [\mathbf{R}_{\max} - \mathbf{R}_{\min}]^{-1} = \Delta\mathbf{R}^{-1} \tag{A9-6}$$

For classification pdfs, we denote their average values as pdf(O|O), pdf(O|C), pdf(C|C), and pdf(C|O), where O stands for the object, C stands for clutter, pdf(O|O) stands for an average value of pdf of an object given object statistics, pdf(O|O) = E{pdf($\mathbf{Y}_n|H$ = object, classification)|O}, and so forth. Combining these average classification pdfs with Eqs. (A9-5) and (A9-6) we define object and clutter average likelihood ratios per pixel, LR_O and LR_C,

$$\begin{aligned} \text{LR}_\text{O} &= r_\text{O} \; \text{pdf(object|object)}/r_\text{C} \; \text{pdf(object|clutter)} \\ &= r_\text{O} \; (2\sigma)^{-1} \; \text{pdf(O|O)}/r_\text{C} \; \Delta R^{-1} \; \text{pdf(O|C)} \\ \text{LR}_\text{C} &= r_\text{C} \; \text{pdf(clutter|clutter)}/r_\text{O} \; \text{pdf(clutter|object)} \\ &= r_\text{C} \; \Delta R^{-1} \; \text{pdf(C|C)}/r_\text{O} \; (2\sigma)^{-1} \; \text{pdf(C|O)} \end{aligned} \tag{A9-7}$$

Expected values of probabilities in Eq. (A9-4) can be expressed in terms of these quantities. According to probability definition, for k = object and n = object,

$$\begin{aligned} f(k|n) &= r_\text{O} \; \text{pdf(object|object)}/[r_\text{O} \; \text{pdf(object|object)} + r_\text{C} \; \text{pdf(object|clutter)}] \\ &= 1/\left(1 + LR_\text{O}^{-1}\right) \end{aligned} \tag{A9-8}$$

And for k = object and n = clutter,

$$f(k|n) = r_{\mathrm{O}}\ \mathrm{pdf}(\mathrm{clutter}|\mathrm{object})/[r_{\mathrm{O}}\ \mathrm{pdf}(\mathrm{clutter}|\mathrm{object}) + r_{\mathrm{C}}\ \mathrm{pdf}(\mathrm{clutter}|\mathrm{clutter})] = 1/(1+\mathrm{LR_C}) \tag{A9-9}$$

The sum in Eq. (A9-4) extends in each frame over the pixels determined by the width of track pdf in the numerators of Eqs. (A9-8) and (A9-9), which is 2σ. The number of items in Eq. (A9-4) for each frame, N_p, is determined by the number of pixels within 2σ, determined by the sampling rate, $N_p = 2\sigma/\tau$, where τ is the size of a pixel. Within these N_p items, there is one containing the object and $(N_p - 1)$ items containing clutter. Thus, the CR bound can be written as

$$\sigma^2_{\mathrm{CR}} \sim \sigma^2\ N_t^{-1} \left\{ \left(1 + \mathrm{LR_O}^{-1}\right)^{-2} + \left(N_p - 1\right) \cdot (1 + \mathrm{LR_C})^{-2} \right\}^{-1} \tag{A9-10}$$

Let us evaluate this expression for a case of a very high signal-to-clutter ratio, S/C $\to \infty$, $\mathrm{LR_O} = \infty$, $\mathrm{LR_C} = \infty$. In this case the expression in parentheses equals 1, leading to Eq. (9.4-8):

$$\sigma_{\mathrm{CR}} \sim \sigma\ N_t^{-1/2}, \quad \text{no clutter} \tag{A9-11}$$

Let us evaluate expression (A9-10) for a case of strong clutter. Consider a very strong clutter case first. Then, statistical distributions of clutter signals are indistinguishable from those of clutter + object signals, therefore, classification pdfs are approximately the same for clutter and clutter + object pixels:

$$\mathrm{pdf}(\mathrm{O}|\mathrm{O}) \sim \mathrm{pdf}(\mathrm{C}|\mathrm{O}) \sim \mathrm{pdf}(\mathrm{O}|\mathrm{C}) \sim \mathrm{pdf}(\mathrm{C}|\mathrm{C}), \quad \text{very strong clutter} \tag{A9-12}$$

Rates r_{O} and r_{C} are determined as follows. One object per ping is expected, $N_{\mathrm{O}} = 1$ for the total pixels per frame $N = \Delta R/\tau$, thus, $r_{\mathrm{O}} = N_{\mathrm{O}}/N \sim \tau/\Delta R$. Clutter is expected with high probability in every pixel, hence $r_{\mathrm{C}} \sim 1$,

$$\begin{aligned} \mathrm{LR_O} &\sim (2\sigma)^{-1}\ (\tau/\Delta R)/\Delta R^{-1} = (2\sigma)^{-1}\ \tau = N_p^{-1} \\ \mathrm{LR_C} &\sim \Delta R^{-1}/(2\sigma)^{-1}\ (\tau/\Delta R) \sim \mathrm{LR_O}^{-1} \end{aligned} \tag{A9-13}$$

Combining Eqs. (A9-10), (A9-12), and (A9-13),

$$\sigma^2_{\mathrm{CR}} \sim \sigma^2\ N_t^{-1}\ N_p^{-1}(1 + N_p)^2 \tag{A9-14}$$

Since $N_p \gg 1$, we obtain Eq. (9.4-9),

$$\sigma^2_{\mathrm{CR}} \sim \sigma^2\ N_t^{-1} N_p, \quad \text{for very strong clutter} \tag{A9-15}$$

The possibility of tracking in strong clutter without additional classification features rests on the fact that clutter is random, while the object is on a track (for linear track say, three or more detections on a straight line would indicate the presence of an object); thus, the CR bound in the limit of very strong clutter is proportional to an error of measuring object position in a single frame, N_p. Hence, the last multiplier in Eq. (A9-15) determines the maximal clutter effect of increasing σ^2_{CR} by a factor of N_p relative to the no clutter case Eq. (A9-11).

Let us consider now finite clutter and utilization of classification features to improve the possible association and tracking performance. Again, one object is expected per frame and clutter is expected in every pixel with some probability, so $r_O \sim \tau/\Delta R$, $r_C \sim 1$, and from Eq. (A9-7) we obtain

$$\begin{aligned} \mathrm{LR_O} &\sim (2\sigma)^{-1}\ \mathrm{pdf(O|O)}/\tau^{-1}\ \mathrm{pdf(O|C)} \sim N_p^{-1}\ \mathrm{LRC_O} \\ \mathrm{LR_C} &\sim \tau^{-1}\ \mathrm{pdf(C|C)}/(2\sigma)^{-1}\ \mathrm{pdf(C|O)} \sim N_p\ \mathrm{LRC_C} \end{aligned} \tag{A9-16}$$

Here, we introduced $\mathrm{LRC_O}$ and $\mathrm{LRC_C}$, the parts of single pixel likelihoods, which are determined by classification features alone.

To obtain qualitative expressions that can be intuitively interpreted in terms of standard detection and tracking procedures, we now split classification feature vectors, $\mathbf{Y}$, into a "detection feature" A, and other classification features, $\mathbf{F}$; $\mathbf{Y} = (A, \mathbf{F})$. (Of course, the CR bound does not depend on any detection procedure; it accounts for concurrent estimation of tracks and association of all pixels with all tracks using all the available information.) For the detection feature, we consider the maximal pixel signal A within N_p pixels (since the object's position on every frame is known at best within N_p pixels). According to the rule of conditional probabilities, single pixel pdfs in Eq. (A9-16) can be written as products of two terms: the pdfs of A and conditional pdfs of other features $\mathbf{F}$:

$$\begin{aligned} &\mathrm{pdf(O|O)} = \mathrm{pdf}(A_\mathrm{O}|\mathrm{O})\ \mathrm{pdf}(\mathbf{F}_\mathrm{O}|O, A); \quad \mathrm{pdf(O|C)} = \mathrm{pdf}(A_\mathrm{O}|\mathrm{C})\ \mathrm{pdf}(\mathbf{F}_\mathrm{O}|\mathrm{C}, A) \\ &\mathrm{pdf(C|C)} = \mathrm{pdf}(A_\mathrm{C}|\mathrm{C})\ \mathrm{pdf}(\mathbf{F}_\mathrm{C}|\mathrm{C}, A); \quad \mathrm{pdf(C|O)} = \mathrm{pdf}(A_\mathrm{C}|\mathrm{O})\ \mathrm{pdf}(\mathbf{F}_\mathrm{C}|\mathrm{O}, A) \end{aligned} \tag{A9-17}$$

Here, $\mathrm{pdf}(A_\mathrm{O}|\mathrm{O})$ is an average value of the object pdf for A distributed according to the object pdf, $\mathrm{pdf}(A_\mathrm{O}|\mathrm{O}) = E\{\mathrm{pdf}(A|\mathrm{O})|\mathrm{O}\}$, etc.

Let us evaluate the above quantities for a case of an exponentially distributed clutter, that is, an average clutter pixel signal has an exponential pdf with variance σ_C,

$$\mathrm{pdf}(A|\text{average clutter pixel}) = 1/\sigma_\mathrm{C}\ \exp(-A/\sigma_\mathrm{C}), \quad A > 0 \tag{A9-18}$$

A maximal object signal is a constant S, so that an object pixel contains object + clutter signal $A = (A_{\text{clutter}} + S)$, distributed according to

$$\mathrm{pdf}(A|\mathrm{O}) = 1/\sigma_\mathrm{C}\ \exp[-(A - S)/\sigma_\mathrm{C}], \quad \mathrm{A > S} \tag{A9-19}$$

The peak clutter pdf can be computed from Eq. (A9-18). We have used the exact distribution for numerical computation of the mean and standard deviation of the peak clutter shown in Table 9.4-1.

For deriving rule-of-thumb expressions, we model the peak clutter distribution with an exponential distribution having same mean and standard deviation,

$$\mathrm{pdf}(A|\mathrm{C}) = 1/\sigma_p\ \exp\left[-\left(A - m_p + \sigma_p\right)/\sigma_p\right], \quad A > m_p - \sigma_p \tag{A9-20}$$

Average values of these pdfs given object or clutter signals are computed as follows:

$$\begin{aligned} \mathrm{pdf}(A_\mathrm{C}|\mathrm{C}) &= E\{\mathrm{pdf}(A|\mathrm{C})|\mathrm{C}\} \\ &= \int_{m_p-\sigma_p}^{\infty} \left\{1/\sigma_p\ \exp\left[-\left(A - m_p + \sigma_p\right)/\sigma_p\right]\right\}^2 dA = 1/2\sigma_p \end{aligned} \tag{A9-21}$$

Similarly,

$$\begin{aligned} \text{pdf}(A_{\text{O}}|\text{O}) &= 1/2\sigma_p, \quad \text{pdf}(A_{\text{O}}|\text{C}) = \text{pdf}(A_{\text{C}}|\text{O}) \\ &= 1/(\sigma_p + \sigma_c)\exp[-(S+B)/\sigma_p] \\ B &= m_c - m_p + \sigma_p - \sigma_c \end{aligned} \tag{A9-22}$$

Combining Eq. (A9-16) through (A9-22),

$$\begin{aligned} \text{LRC}_{\text{O}} &= (\sigma_p + \sigma_c)/2\sigma_c \ \exp[(S+B)/\sigma_p] \cdot \text{LRCF}_{\text{O}} \\ \text{LRC}_{\text{C}} &= (\sigma_p + \sigma_c)/2\sigma_p \ \exp[(S+B)/\sigma_p] \cdot \text{LRCF}_{\text{C}} \end{aligned} \tag{A9-23}$$

Combining Eqs. (A9-10), and (A9-16) through (A9-23),

$$\begin{aligned} \sigma_{\text{CR}}^2 \sim \sigma^2 \ N_t^{-1} \Big\{ & \Big[1 + N_p \ \left[\left(\sigma_p + \sigma_c\right)/2\sigma_c\right]^{-1} \\ & \exp\left[-(S+B)/\sigma_p\right] \cdot \text{LRCF}_{\text{O}}{}^{-1}\Big]^{-2} \\ & + (N_p - 1) \cdot \left[1 + N_p \left(\sigma_p + \sigma_c\right)/2\sigma_p \right. \\ & \left. \exp\left[(S+B)/\sigma_p\right] \cdot \text{LRCF}_{\text{C}}\right]^{-2} \Big\}^{-1} \end{aligned} \tag{A9-24}$$

The above expression can be simplified for $\text{S}/\sigma_c \lesssim 2$, and accounting for $N_p \gg 1$, leads to Eq. (9.4–14).

NOTES

1. Note a subtlety in notations: boldface is used to denote vectors and matrices (e.g., **IM**); their components are denoted without bold (e.g., IM^{ab}); but, in $(\mathbf{IM}^{-1})^{ab}$, the boldface is used to emphasize that the matrix **IM** is inverted and indexes a, b refer to the elements of the inverted matrix.
2. The difficulty of obtaining CR bounds for tracking in clutter is due to a need for performing association between sensor measurements and objects. When tracking multiple objects, or tracking in clutter, the task is usually divided into two subtasks or functions: of associating sensor measurements with individual objects (or other signal sources, such as noise and clutter; the association function) and another one of estimating tracks (positions, velocities, etc.) of the objects (tracking function). Doubts have been expressed that CR bounds accounting for both functions could be derived in principle (Daum, 1990). In Daum (1990), an approach to quantifying performance bounds was derived assuming multiple hypothesis tracking (MHT) association. Formulation of bounds independent of the association algorithm was considered in Graham and Streit (1994). Explicit expressions, however, were not obtained, and the difficulty of the problem of formulating the bounds was analyzed. Streit (1995) obtained performance bounds in terms of accuracy of an initiated track assuming the nearest neighbor association. Assumptions of a particular association algorithm had to be made in these publications in order to make the problem amenable to the analysis. The nearest neighbor association assumption is appropriate for a case of weak clutter. The number of hypotheses that should be evaluated in MHT association grows

exponentially in heavy clutter, thus practical utility of bounds based on MHT association should also be expected within the area of relatively weak clutter. The general association and estimation CRB are described in this chapter following Perlovsky (1989a, 1992a, 1997a). These derivations were made possible by the development of adaptive fuzzy AZ-logic described in this book.

BIBLIOGRAPHICAL NOTES

The Cramer–Rao bound (CRB), and CR theory (Cramer, 1946).

The general association and estimation CRB are described in this chapter following Perlovsky (1988c, 1989a, 1992a, 1997a).

Discussions of the Cramer–Rao bound for tracking (Daum, 1990; Graham and Streit, 1994; Streit, 1995).

PROBLEMS

9.1–1 Prove Eq. (9.1-11): $\text{cov}\{\hat{\mathbf{S}}, (LL_{;\mathbf{S}})^{\mathrm{T}}\} = 1$. *Hints:*

1. Consider the definition, $\text{cov}\{\hat{\mathbf{S}}, (\partial/\partial\mathbf{S})\ \ln \text{pdf}(\mathbf{x})\} = E\{[\hat{\mathbf{S}} - E\{\hat{\mathbf{S}}\}] \cdot [(\partial/\partial\mathbf{S})\ \ln \text{pdf}(\mathbf{x}) - E\{(\partial/\partial\mathbf{S})\ \ln \text{pdf}(\mathbf{x})\}]\}$.
2. Note that $E\{E\{\hat{\mathbf{S}}\}(\partial/\partial\mathbf{S})\ \ln \text{pdf}(\mathbf{x})\} = E\{\hat{\mathbf{S}}\}E\{(\partial/\partial\mathbf{S})\ \ln \text{pdf}(\mathbf{x})\}$, and use (9.1-9).
3. Obtain $\text{cov}\{\hat{\mathbf{S}}, (\partial/\partial\mathbf{S})\ \ln \text{pdf}(\mathbf{x})\} = E\{\hat{\mathbf{S}}\ (\partial/\partial\mathbf{S})\ \ln \text{pdf}(\mathbf{x})\}$, and use (9.1-7).

9.1–2 Prove Eq. (9.1-16). *Hints:* Prove that the variance of a sum of independent variables equals a sum of individual variances. Then notice that expected values of all LL_n are the same.

9.1–3 Prove Eq. (9.1-18) and obtain CRB for the mean of a Gaussian distribution. *Hints:*

1. Take the derivative of the Gaussian distribution Eq. (1.3-9),

$$LL_{1;\mathbf{M}} = (\partial/\partial\mathbf{M})\ \ln\ G(\mathbf{x}|\mathbf{M}, \mathbf{C}) = (\partial/\partial\mathbf{M})(-0.5\ \mathbf{D}^{\mathrm{T}}\ \mathbf{C}^{-1}\mathbf{D}) = (\mathbf{x} - \mathbf{M})^{\mathrm{T}}\ \mathbf{C}^{-1}$$

2. Evaluate $\text{var}\{LL_{1;\mathbf{M}}\} = E\{LL_{1;\mathbf{M}} \cdot (LL_{1;\mathbf{M}})^{\mathrm{T}}\} = E\{\mathbf{C}^{-1}(\mathbf{x} - \mathbf{M})(\mathbf{x} - \mathbf{M})^{\mathrm{T}}\ \mathbf{C}^{-1}\} = \mathbf{C}^{-1}E\{(\mathbf{x} - \mathbf{M})^{\mathrm{T}}(\mathbf{x} - \mathbf{M})\}\mathbf{C}^{-1} = \mathbf{C}^{-1}\mathbf{C}\mathbf{C}^{-1} = \mathbf{C}^{-1}$.

9.3–1 Derive Eq. (9.1-3) from Eq. (9.3-13).

9.4–1 Derive Eq. (9.4-3). *Hints:*

1. Use index notations for the model $M_k^i = R_k^i + V_k^i t_n$, where index $i = 1, 2$ corresponds to the coordinates (range and angles, or two angles), and for deviations, $D_{nk}^i = \left(x_n^i - R_k^i - V_k^i t_n\right)$.
2. Use mixed index-vector notations for the parameters, $a = i$, $S_k^a = \left(R_k^a, V_k^a\right)$.
3. Derive the model derivatives

$$M_k^{i;a} = \partial M_k^i/\partial S_k^a = \delta^{i,a}\left(\partial/\partial S_k^i\right)\ \left(R_k^i + V_k^i t_n\right) = \delta^{i,a}(1, t_n)$$

where $\delta^{i,a} = 1$ for $a = i$, and 0 otherwise.

4. Substitute these derivatives into (9.1-23), and obtain the information matrix in mixed index-matrix notations [verify by components that position-velocity components are given by the matrix $(1, t_n)\ (1, t_n)^{\mathrm{T}}$]:

$$\mathbf{IM}^{ab} = \sum_n E\left\{ f^2(k|n)\ D_{nk}^{i'} D_{nk}^{j'} \right\} \left(\mathbf{C}_k^{-1}\right)^{i'i} \left(\mathbf{C}_k^{-1}\right)^{j'j} \delta^{i,a}(1, t_n)\delta^{i,b}(1, t_n)^{\mathrm{T}}$$

$$= \sum_n E\left\{ f^2(k|n)\ D_{nk}^{i'} D_{nk}^{j'} \right\} \left(\mathbf{C}_k^{-1}\right)^{i'a} \left(\mathbf{C}_k^{-1}\right)^{j'b} (1, t_n)(1, t_n)^{\mathrm{T}}$$

In matrix notations, this expression is written in Eq. (9.4-3).

chapter 10

INTELLIGENT SYSTEM ORGANIZATION
MFT, Genetic Algorithms, and Kant

The first philosophical system embracing multiple aspects of human mind in their interaction was developed by Kant. His philosophy was a turning point in the entire history of human thought and is considered a beginning of scientific psychology. Kant developed a rational theory of intelligence by establishing that the world of phenomena depends on mind. We overview the main components of intelligence identified by Kant: Understanding, Judgment, and Reason, and relate them to the basic MFT components: internal models, similarity measures, and adaptation. It turns out that the intelligent agents of MFT implement the process of thought described by Kant. We discuss the heterohierarchical organization of the mind and relate it to the Kantian problem of synthetic judgments a priori.

Emotions and perception of beauty are fundamental to the human mind, alike in everyday life, arts, and sciences. Still, the concept of beauty is mystifying. The first step toward mathematics of beauty is made in this chapter. It is founded on the relationship between MFT and the Kantian theory of mind. MFT's instinct-will for learning, according to Kant, is a basis for emotional intellectual abilities, for the beautiful and sublime. Future development of MFT will include complex internal models addressing conscious and unconscious aspects of mind. We overview the mathematics of learning complicated structural models: genetic algorithms, complex adaptive systems, and semiotics, and relate them to MFT. The MFT intelligent agents are shown to implement mathematically the dynamic loops of semiosis; they are "vortexes" of symbol formation, vortexes of thought.

Kant overturned the understanding of the relationship between the mind and the world by considering the specific a priori contents of mind that enable its functioning. The philosophy of Pure Spirit came close to the scientific method. He developed a rational explanation of mind as a system, if not in its entirety, still in its most interesting "higher" intellectual abilities. Many aspects of Kant's theory were further developed by a number of philosophers and psychologists, including Schopenhauer, Hegel, Nietzsche, Freud, Jung, Bergson, and Jaspers. Still, the original theory of Kant remains unsurpassed in its comprehensive treatment of mind as a system. And mathematical theories of intellect have remained far removed from the penetrating depth of his understanding and are inadequate for coming even close to the width of his analysis. This does not have to be so, for Kant's analysis is rational and therefore can perfectly serve as a foundation for developing the mathematical

theory of mind. A first step toward rectifying this deficiency of the mathematical theories of intellect is undertaken in this chapter.

10.1 Kant, MFT, and Intelligent Systems

What is intelligence? It is still shrouded in mystery. A lot is understood, but much still is unknown and, most likely, will remain unknown for a while. Intelligence is attributed to natural and artificial systems and some of these systems are very simple, whereas others are very complex. At lower levels, intelligence is an ability to sense the environment and to control the body (machinery) toward achieving a few predetermined goals. At higher levels intelligence includes abilities of thinking, including recognition and formation of concepts, developing complicated internal representations of the outer world and self, understanding, and language ability; planning behavior, including direction of attention, definition of goals and subgoals, and the ways to achieve them; acting within itself and in the outer world; an ability of judgment, including feeling and emotions; and abilities of intuition, learning, consciousness, creativity, and a mysterious feeling of freedom of will.

In three volumes on the *Critique of Pure Reason*, *Critique of Judgment*, and *Critique of Practical Reason*, Kant explained a wide variety of intellectual experiences based on three fundamental abilities or faculties of mind: Understanding, Judgment, and Reason. Each is based on specific a priori principles or instincts contained in the mind: concepts, correspondence between concepts and manifold of sensory data, and will or desire. Understanding is a faculty of concepts, a source of general notions. Judgment is an ability to see that a particular case comes under the general rule. And Reason is an ability to draw conclusions in terms of generating behavior. (The most intellectually important type of behavior, interwoven with higher intellectual abilities and emotions, Kant considered to be the behavior of learning.) In this chapter, Understanding, Judgment, and Reason are capitalized when they refer to the fundamental abilities of the mind, or, alternatively, to modules of an intelligent system. These three abilities correspond to the three aspects of consciousness: knowledge (of concepts), feeling (of correspondence between concepts and outer world), and desire (to act).

Even though Kant devoted a separate book to each ability, they should be combined within a dynamic system constantly exercising all three abilities in their interaction. This chapter takes a step toward considering intelligence as an interacting system. We will see that even relatively simple MFT paradigms considered in previous chapters contain seeds of mathematical modeling of the three main elements of intelligence identified by Kant. MFT carries Kantian analysis further: it is a dynamic system in which the three abilities identified by Kant exist in the process of constant interaction, as it were in a "vortex." This vortex describes learning of a concept as a dynamic formation of a symbol. We overview some of the higher intellectual abilities, along with attempts at their rational explanation and mathematical description.

10.1.1 Understanding Is Based on Internal Models

An internal model is a basis of intelligence. Even at the lower levels, say, of a lobster sensing and grabbing food, with the axons of sensing cells "wired" directly into the neurons that

control muscles, we can talk about internal model. Because the signal that a "food-sensing" neuron sends to a "muscle-neuron" indicates an internal lobster's representation of food. There is no such thing as "food" in the ocean; "food" is a dynamic process of interaction between an object in the ocean, sensing neuron (that forms an internal representation-signal), grabbing neuron, and other relevant neural aspects of the lobster's experience. A lobster's mind has literally few neurons, and if our final goal would be modeling of a lobster's mind, we will directly proceed to studying its wiring diagram without such nebulous and not obviously useful concepts as a lobster's internal models.

Our aim, however, is to understand and model higher levels of intelligence. At higher levels, a complete "wiring diagram" of a neural system, even if available, would be so complicated that it does not furnish an understanding of the basic principles of mind. A significant part of the brain is involved with internal models (storing, updating, and using them). Our ability to recognize concepts, even simple ones, such as objects, is due to internal models or representations of concepts. Understanding, first, consists of concepts in our mind along with their interrelationships. Higher levels of understanding, such as understanding of meaning, involve a complex internal model composed of a large number of submodels-concepts with multiple interconnections among them. Possibly, every particular phenomenon of understanding–meaning exists only within a limited domain, within a certain situation, or with regard to a certain goal. Meaning of a concept is then modeled by including this concept within a set of situationally relevant other concepts and goals. Meaning requires a hierarchical system: the understanding of the meaning of a concept requires a point of view from the next levels in a hierarchy, above the level of the concept's inner model and its recognition. Thus, the meaning of a lower level concept is included into a higher level concept. However, relationships among levels are not rigidly fixed: formation of certain concepts involves multiple hierarchical levels, and the relative position of concepts in the hierarchical levels might be situationally dependent. Thus, heterarchical hierarchy might be a better term. Explanation of mind as based on a priori inner models ascends to Plato and Aristotle. Kant identified a priori inner models as a separate faculty of mind that he called Understanding. The mind's operations with a priori concepts Kant calls the domain of Pure Reason.

The main question that the analysis of Pure Reason shall answer, according to Kant, is "How are *synthetic judgments a priori* possible?" Here, *synthetic judgments* are conclusions or statements that derive a new meaning by combining several concepts. For example, "a horse has four legs" is not a synthetic statement, because the concept "horse" assumes four legs (it is an *analytic* statement, a consequence of the definition of "horse"). Another example: "a car has four wheels." This is a synthetic statement, because four wheels are not a necessary part of the definition of "car" (one can build a three-wheel car); however, this is *not an a priori* statement, because the truth of this statement depends on experience (with particular cars) and is not a universal truth. The first laws of physics and most theorems of mathematics, according to Kant, are synthetic judgments a priori. They are not tautologies, because they contain new information. They are *a priori,* because their *universal* validity cannot be based merely on experience, but requires something else. This "something else" Kant identified as a specific a priori faculty of pure reason. In our theory of mind, this specific faculty is represented in hierarchical models: next levels in a hierarchy contain synthesis of the lower level concepts. This synthesis is of an a priori nature, because the hierarchical structure of the internal model is of an a priori origin. Thus, development of

hierarchical models is a key to mathematical modeling of Understanding and Pure Reason. Making this hierarchy adaptable and situationally dependent is an additional challenge.

For example, in a tracking MFT system considered in Chapter 7, concepts or categories are given by a priori models of moving objects of interest (targets), noise, and other objects of no interest (clutter). Moving objects are characterized by velocity, which is an integral part of their model; so a statement that moving objects have a property of velocity, according to Kant, is an *analytic judgment* (not a synthetic one). But when we make a decision that a certain object is a target, this decision is based on comparing the target-clutter likelihood ratio to a threshold; this procedure is not an integral part of the moving-object concept, thus, it is not a tautology, it is a *synthetic judgment.* Because this procedure is universal (it is always used for this purpose), it is *an a priori judgment*. This procedure is an example of a synthetic judgment a priori. Mathematically, this procedure is contained in the next hierarchical level of MFT (above the level of individual objects). This example illustrates that an ability for synthetic judgments a priori is due to a hierarchical organization of an intelligent system.

Explanation and modeling of the phenomena of meaning and understanding require also including them within behavior generation and acting of an intelligent system. The acting could be inside, within an intelligent system, or outside of the intelligent system, into the outer world. Actions, corresponding to the goal or situation (internal or external), constitute a part of meaning. There is also another aspect of acting out in the external world noted by Freeman, who introduced a concept of external representations in the world. Our external acts and their results (being perceived by ourselves and others) from gestures and utterances to our entire culture, as it exists in the outside world, are external representations. To the extent that external representations are included in the Kantian cycle of concept formation, they can be viewed as parts or extensions of our internal models. Computer simulations are a perfect example of such extensions of internal models. The entire culture is an external representation of concepts of mind.

According to Kant, logic gives laws of understanding, or laws of relationships among a priori concepts. Here, in the world of Ideas, there is a significant domain of applicability of Aristotelian logic. For example, an internal model-concept of an object is either that of target or not, according to the Aristotelian logic law of excluded third (and this logic is different from fuzzy logic of judging which real signal belongs to which concept). This domain of Aristotelian logic encompasses nonadaptive aspects of the a priori models. Kant missed a need for adaptation and he did not notice the Aristotelian emphasis on the changing nature of Forms in the process of transformation from potentiality to actuality. To the extent that the a priori models can adapt, they are fluid, noncrisp, fuzzy. Development of adaptive hierarchies of models is a challenge for future research.

To summarize the MFT relationship to Kant's Understanding: MFT ability for Understanding or forming concepts is due to a priori internal models, and an ability for synthetic judgments a priori is due to an a priori hierarchy of models and relationships among them.

10.1.2 Judgment Is Based on Similarity Measures

For internal models or concepts to be useful, there should be a way of relating them to experience. In other words, we should be able to recognize individual phenomena according to general concepts and to decide which aspect of the empirical world corresponds to which

concept. Kant called this ability Judgment and considered it one of the three main abilities of the mind. Judgment is an ability to see that a particular case comes under the general rule. In our MFT theory, Judgment is mathematically represented by similarity measures. MFT contains a measure of similarity between the internal model and the world, as well as between each submodel-concept and a particular subset of sensory data. This is an a priori property of MFT and, according to Kantian analysis it is an a priori property of our mind.

Kant differentiates determinant and reflective aspects of Judgment. Finding particular subsets of sensory data corresponding to a specified concept-submodel is the determinant Judgment. And finding the concept corresponding to the data is the reflective Judgment. MFT contains mechanisms of both, determinant and reflective Judgment. Within the iterative loop of MFT adaptation, determinant Judgment is given by the association (segmentation) of data with concept-submodels, and reflective Judgment is given by selecting the concept-submodel most similar to a particular subset of the data. In MFT the determinant Judgment is given by the fuzzy associations, $f(k|n)$, with n designating data and k designating the concept-submodels; and the reflective Judgment is given by the fuzzy associations, $f(k|n)$ and adaptation laws. These relationships can be summarized as

Kant's determinant Judgment:	$M \to X,$	MFT: $\mathrm{l}(X\|M)$
Kant's reflective Judgment:	$X \to M,$	MFT: $\mathrm{l}(X\|M)$ + adaptation

Why is Nature in its manifold knowable to our mind? Is it due to a specific property of Nature or to a specific property of mind? In other words, what makes it possible for our Understanding and Judgment to function in the way described above? Kant's answer is that this possibility is due to a special a priori property of our mind. This property is the purposiveness of our internal representations (models). Understanding and Judgment are so constructed that internal representations of empirical events and objects appear to us as purposive (the purpose includes first, a correspondence between our internal representations and the world, and second, an ability to learn or to improve this correspondence). This purposiveness provides a foundation for the development of higher faculties of mind including higher emotions, and the notions of the beautiful and sublime.

A reader might wonder if this discussion is too philosophical and irrelevant to a mathematical theory of mind? The relevancy of this discussion is in that it guides us in constructing internal models, measures of similarity, and in developing evolutionary theories explaining these abilities. The models and similarities are constructed so that they have a purpose or meaning within the intelligent system, which is the mathematical description of the intentionality of the intellect. This intentionality includes the correspondence to the world and adaptivity that provides for learning. And it is needed so that the "lower level" instincts for survival, for performing specific tasks, etc. can be more efficiently satisfied (by a living being or a robot). Intentionality provides a background for a mathematical theory of higher faculties of mind, including the possibility for mathematical treatment of the beautiful and sublime. And an evolutionary theory must lead to these abilities.

To summarize the MFT relationship to Kant's Judgment: MFT ability for Judgment is due to similarity measures and fuzzy concept memberships, which select data corresponding to the concepts of Understanding and select concepts corresponding to the data, in every cycle of the iterative MFT loops.

10.1.3 Reason Is Based on Similarity Maximization

Judgment mediates between concepts of Understanding and concepts of Reason (will, and freedom). In particular, reflective Judgment ascends from the particular to the universal, from sensory data to concepts. Its principle is an ability to learn, which Kant called the purposiveness of intellect toward the object. This ascendance from data to concepts is practically realized by Reason. In MFT functioning, finding a submodel corresponding to a piece of data (Judgment) is followed by adaptive modification of the model, which is the act of will according to the learning principle (law) of Reason. Reason provides laws for behavior, and MFT paradigms considered in previous chapters were concerned with one type of behavior: learning behavior as adaptation of the internal model. Modification of models in MFT is governed by the principle of maximum similarity between the model and data. The MFT parameter-adaptation equations maximizing the similarity give the laws of Reason. Thus, MFT provides for a mathematical description of a will for learning, a will for improvement of its internal representations of the world and the laws of Reason governing this will.

Kant emphasized a fundamental nature of the antinomy between causality and freedom and severely criticized philosophers, who underestimated the difficulty of the causality–freedom antinomy. And, his criticism still applies to a researcher who is too cavalier about resolving this antinomy.[1] The fundamental source of difficulty is in that freedom is an opposite of randomness. Freedom supposes causality. If there is no causality, there could be no freedom. But if the world's laws are causal, how could freedom be explained? Kant made a step toward resolving this antinomy. He assigned the concept of causality to Understanding, where causality is an a priori concept of understanding the nature, the world of phenomena. And he assigned the concept of freedom to Reason, where it is an a priori concept governing human desire and will. According to Kant, freedom belongs to a noumenal world; it originates from the unknowable nature of a human-in-itself. A next step toward resolving this antinomy should be attempted by identifying the unknowable human-in-itself with our unconscious and developing a physical theory of conscious and unconscious aspects of mind.

10.1.4 Hierarchical Organization of Intelligent Systems

Let us summarize the discussion of the previous three sections. The three Kantian abilities are organized in a dynamic loop of MFT as illustrated in Fig. 10.1-1. The MFT loop maintains the current situational awareness that is the correspondence between the internal model and the world. This includes updating the parameters of a large number of learned concepts (or recognized objects), recognizing new objects-concepts that might appear in the input data, terminating old concepts that are not relevant any longer, and searching in the data for particular concepts or objects of interest (say, food). Association and parameter estimation for each submodel (corresponding to an object-concept) is computed in parallel with other submodels. Therefore, the MFT association–adaptation loop consists of a large number of individual concept-loops, or intelligent agents.

Each concept-loop is an "agent": it has its own rules for activation, performance, interaction with other concept-loops, and termination. Some concept-agents may interact with each other [if for some data n, $f(k|n)$ are nonzero for several k-models], and many

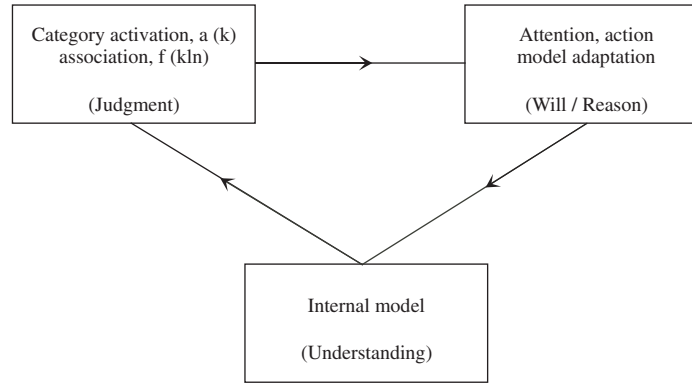


Figure 10.1-1 Internal cycle of intelligent system operation. Kant vs. MFT.

are independent from each other to a significant extent. The concept-agents for learned concepts, the parameters of which are not expected to change too fast, might be updated sparingly, less often than newly activated agents. Agents might be initiated by a bottom-up message (new data input from sensors, or lower levels of data processing), top-down message (upper level activating a new submodel-agent, if it is decided that the currently active submodels are inadequate, or if the system goal is changed), or heterarchical type message [when two agents of a similar type are competing for the explanation of the same piece of data, one agent might send a terminating (apaptosis) message to another]. Both, bottom-up and top-down flow of signals may continue all the time, so that the loop is permanently active.

Each active agent continuously exercises a sequence of the three Kantian faculties: computes a submodel (Understanding), evaluates a similarity measure and fuzzy membership (Judgment), and acts by changing model parameters, or by sending behavior signals to the actuators acting in the outer world or to a lower processing level (Reason). Actions lead to changes in the input data by producing changes observed by sensors and by redirecting sensors. This reinitiates the loop of MFT. Specific motor mechanisms are used by animals and humans for efficient movements of various parts of the body by coherent control of multiple groups of muscles. Motor mechanisms are a subject of specialized literature, and in this book we do not discuss them. Although it is worth emphasizing that the mechanisms of bodily motions are an integral part of our perceptions. We perceive with all our body, therefore motor mechanisms are parts of the Forms of mind (Lakoff and Johnson, 1983). Intellectually, the most important types of actions, according to Kant, are those related to learning, and in Chapter 7, we touched on attention mechanism related to the action of learning.

An intelligent system contains hierarchies and heterarchies of multiple levels of the Kant–MFT loops discussed above. Within a hierarchy, every next level exercises syntheses of lower level concepts based on synthetic a priori models at each level. A hierarchical MFT structure is illustrated in Fig. 10.1-2. At the lower level, the input data are provided by sensors and acting is exercised in the world. At higher levels, the input data are provided by the concepts activated at lower levels and actions are signals sent to a lower level. In addition, every agent exercises actions of learning by adapting its own models.

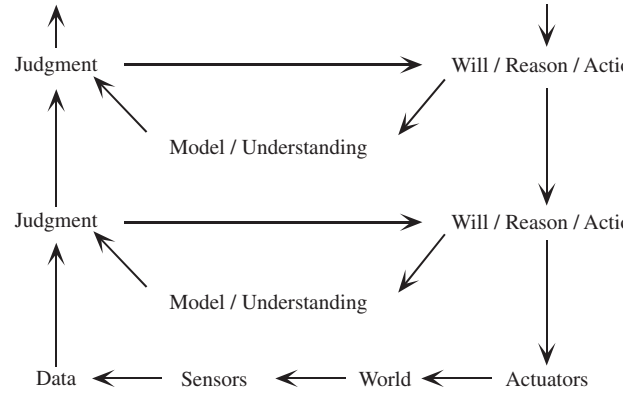


Figure 10.1-2 Hierarchical organization of Kant–MFT intelligent system.

More flexible heterohierarchies are organized as follows. Agents post their output messages on a "web page" or bus-type connections, where it is available to those other agents, who have learned (evolved) to use this type of information. The messages contain, in addition to the activation degree, an identifier information (which agent, or what type of agent sends the message, and estimated parameters of models). Of course, every agent also learns by adapting its own models.

Neural connections in the brain indicate the existence of both types of architectures: vertically connected hierarchies and horizontally connected bus-type heterarchies (Fig. 10.1-3). Horizontal interactions are an important part of the functioning of MFT agents (submodels); see, for example, a competition layer of MLANS architecture in Fig. 4.3-2. This architecture was also essential in Chapter 7 for object-search agents: these agents are activated by incoming data and have a significant intralevel interaction, including exchange of apaptosis (termination)-type messages, in a case in which two agents are tracking the same object.

10.1.5 Aristotle, Kant, Zadeh, MFT, Anaconda, and Frog

There is a folktale about an Anaconda that can hypnotize a Frog so that it jumps into the Anaconda's mouth. Let us analyze this tale within the context of philosophical and mathematical concepts of mind. The Frog's mind is very simple (and we will even further simplify it for our conceptual analysis). There are hardwired connections between the receptive cells in the Frog's eyes and certain muscle groups in its legs. These connections form the Frog's mind internal model, illustrated in Fig. 10.1-4. When there is a small moving object nearby, the Frog jumps toward it and eats it. When there is a big moving object nearby, the Frog jumps away and flees. The Frog's faculty of Understanding is given by three a priori concepts: moving, big, small. There is very limited fuzziness and adaptivity to these concepts, just an estimation of the exact position and speed of moving objects. But the Frog does not analyze in great detail the entire scene in front of him. The Anaconda's mind coevolved with the Frog's mind, so that its predatory concepts of Understanding match the deficiency of the Frog's. The Anaconda can stay still for hours, so that the Frog does not

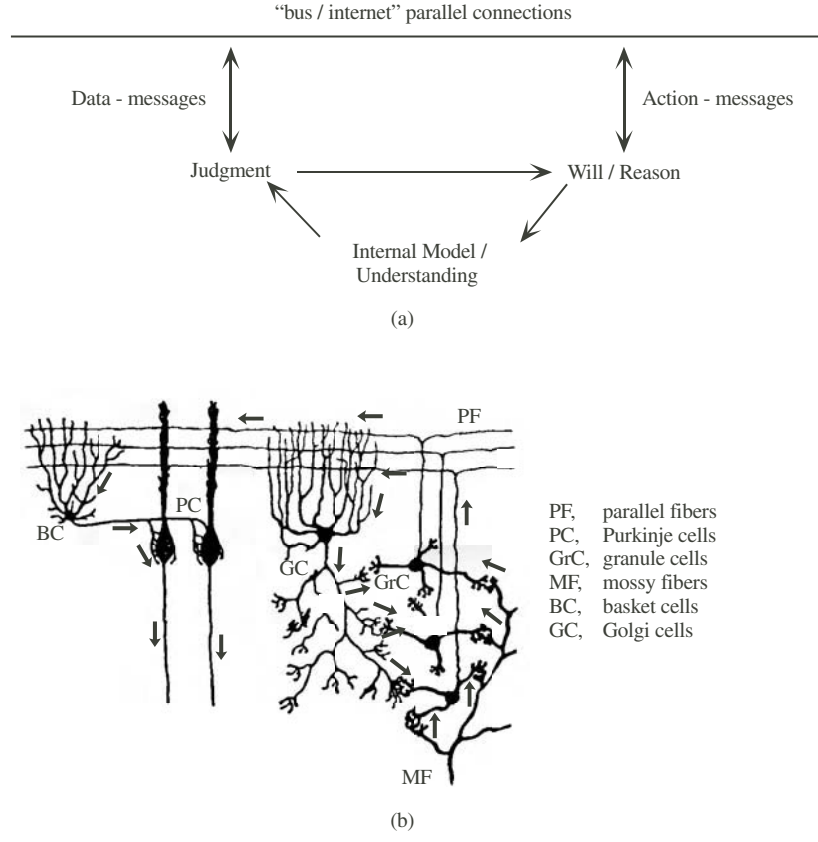


Figure 10.1-3 Heterohierarchical organization; (**a**) Kant–MFT intelligent system; (**b**) systems of the cerebellar cortex.

recognize it as a threat. When the Frog is nearby, the Anaconda just moves its eye, and the Frog perceives the eye as a small moving object, so it jumps right at it.

Now, let us compare the Frog's internal model in Fig. 10.1-5a to the concepts of philosophers and mathematicians about the mind. At the top of the hierarchy of the Frog's mind, there is a most general concept available to it: the totality of the universe. It is composed of two concepts: desirable and nondesirable. This is the dichotomy of Aristotelian logic. This dichotomy is a synthetic judgment a priori: it is synthetic in that it is composed of the two concepts (prey, predator), and it is a priori, because it is a universal truth for the Frog's mind (and all Frogs always use it). The concepts of prey and predator, again, are the a priori concepts of the Frog's Understanding faculty. And, again, they are synthetic judgments a priori: each is an a priori synthesis of the two a priori concepts that are lower in the hierarchy. A lower hierarchical level of the Frog's mind internal model is similarly organized; it is composed of nonfuzzy, Aristotelian logic type a priori concepts (small, big,

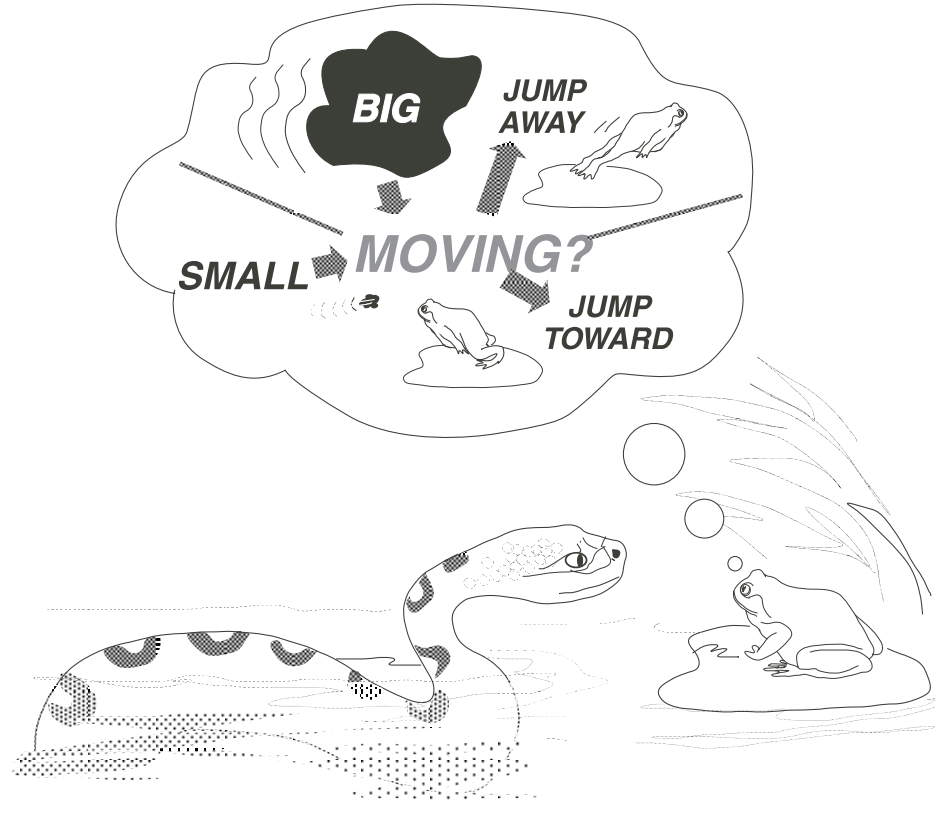


Figure 10.1-4 An artist's illustration of the frog's faculty of understanding and its interaction with anaconda.

moving), which are synthetic judgments a priori. At the very bottom of the hierarchy, there are the receptive fields of the Frog's eyes. These have some degree of adaptivity and a necessary degree of fuzziness: at least, they can adapt to the actual position and speed of the moving objects. But, they are not fuzzy enough and adaptive enough to adapt to the Anaconda's trick. Figure 10.1-5b illustrates the Kant–MFT loop of the Frog's mind.

There is a necessary connection between fuzziness and adaptivity. Crisp, nonfuzzy concepts of understanding do not "see" anything that does not fit them perfectly. Therefore, they cannot learn through gradual adaptation. Adaptation requires fuzzy concepts of understanding, that can be associated (by Judgment) with a variety of conditions so that adaptation to varying conditions is possible. Aristotelian logic-concepts cannot learn gradually, but can possibly evolve due to evolution, if a majority of nonadaptive Aristotelian logic frogs will be eaten up by anacondas. Maybe something like this happened in some isolated frog communities several hundred millions years ago, leading to the evolution of

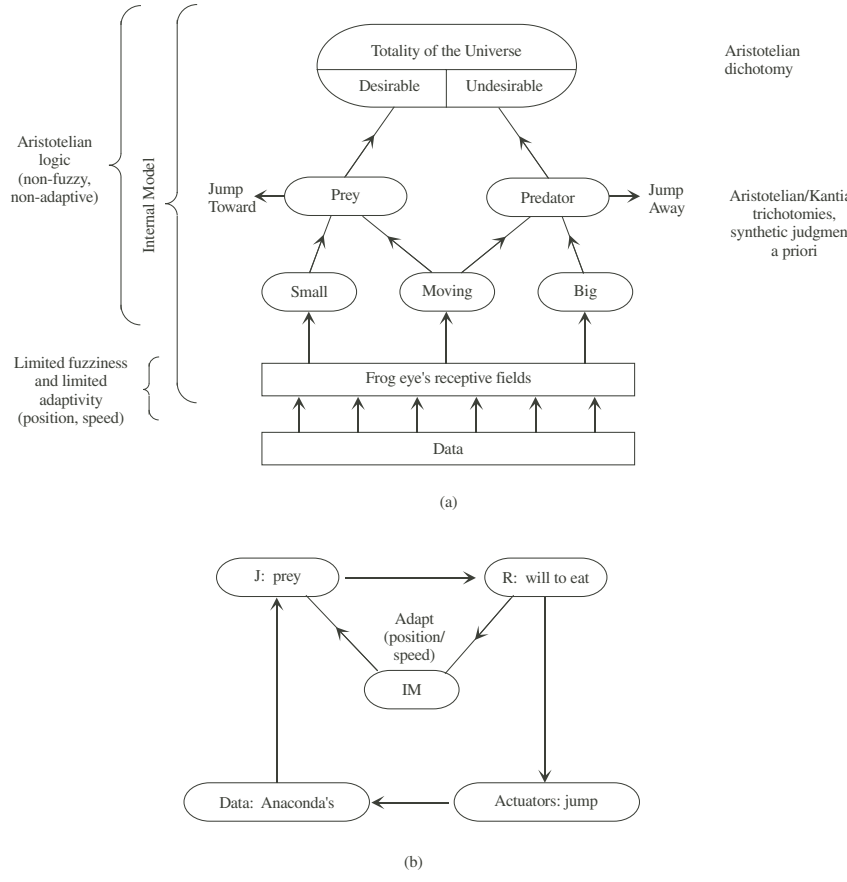


Figure 10.1-5 Philosophy and mathematics of frog eaten by anaconda; (**a**) frog's mind internal mode; (**b**) Kant–MFT operation loop of frog's mind.

more complicated animals, whose understanding uses adaptive Aristotelian Forms governed by nonAristotelian fuzzy logic.

10.2 Emotional Machine (Toward Mathematics of Beauty)

10.2.1 Cyberaesthetics or Intellectual Emotions

If you pinch your finger, it hurts, and an ability to feel the pain is obviously an a priori faculty, which is necessary for survival to such an extent that it is shared within the entire animal kingdom. This "lower" origin of feelings separates them from our higher cognitive abilities. And there is a long-standing line of thought that separates and contraposes feelings and thinking, emotions and intellect. But in 1787 in a letter to his friend, Kant wrote that he had discovered a new type of a priori principle, the feelings of pleasure and pain, which

he found to be a necessary part of our intellect. Kant came to the conclusion that Judgment is based on the feeling of pleasure caused by the harmony or correspondence between our internal representations-concepts and empirical phenomena. The new principle governs "intellectual emotions." These "higher" emotions are not separated from thoughts, but they are combined together in a dynamic process of Kant–MFT agent-loops. Mathematical apparatus describing higher emotions in MFT is given by similarity measures. And a thought process is a loop, a vortex of concepts, emotions, and adaptation-learning actions.

Related to higher emotions is an ability to perceive beauty, which is a universal and fundamental property of the human mind. It is important not only in the field of fine art, but it pervades human experience. Ancient pottery and cave-wall paintings demonstrate the primordial origin of aesthetic emotions. There are well-known statements by famous scientists explaining that the first and foremost test of a scientific theory is its beauty. But mathematical attempts to model mind, so far, have not touched the subject of beauty, and the directions along which this could be attempted seem to be hidden in mystery and not accessible to scientific investigations. Here, I attempt a first step in this direction.[2] On the road to the future science of cyberaesthetics, I have found that I am primarily assisted by Kant, who with scrupulous detail analyzed the rational mechanisms of beauty and other higher emotional faculties.

10.2.2 Purposiveness, Beauty, and Mathematics

When designing an intelligent system, for example, a robot, we decide what kinds of objects the robot should be able to recognize and we supply the robot with the internal models of these objects. From the robot's perspective, only those objects exist that it can recognize. Every object has a purpose of being recognized (in addition to any other purpose the robot may put this object for). A universal purpose of any object is its concept; for an object to have any purpose for a particular intelligent system, the object's concept has to exist in the system. This is a design principle of any intelligent system. This design principle is applicable to us: evolution (or God) designed us so that we can find our way around those objects that we recognize in nature. The basic principle of design is that nature appears to us as having a purpose. The purposiveness of nature is the a priori part of our representations and it harmonizes nature with our desire for knowledge and produces the feeling of pleasure (or pain, if chaos is encountered[3]).

Knowledge about objects comes from experience and from the a priori concepts (Understanding). The role of Judgment in this process was discussed in the previous section: it is an objective, or cognitive aspect of Judgment. In this section, we concentrate on the subjective aspect of Judgment, which relates to the subject and not to the object. This subjective aspect is the satisfaction and feeling of pleasure that is bound with the harmony between our internal representations and an object. Kant calls this the *aesthetic* aspect of Judgment; it relates more to emotions than to cognition (even though all aspects are combined in every act of perception and cognition). This aesthetic aspect of Judgment is related to the "pure" purposiveness of our representations, which is separate from any specific purpose that an object can be used for, and includes only the knowledge itself. Thus, android-robots capable of learning have to be designed so that they have an aesthetic affinity to knowledge. In MFT this is given by the similarity, $l(\mathbf{X}|\mathbf{M})$, that relates a particular case $\mathbf{X}$ to the general concept $\mathbf{M}$, without any further specific purpose the object-$\mathbf{X}$ can be used for.

To the extent that the purposiveness is felt in its pure form and is bound to its a priori nature, the object is called beautiful. The nature of beauty is related to an interest not in the object, but in the subject: what I make out of this representation *in myself*. Beautiful is what coincides with the purpose of acquiring more knowledge and improving the harmony between the internal representations and Nature. Kant discusses two higher intellectual aesthetic abilities: feelings of the beautiful and the sublime. Beautiful involves the relationship between Judgment and Understanding and sublime involves the relationship between Judgment and Reason. The feeling of the sublime moves the Reason to act toward improvements of internal representations. In other words, the beautiful involves the aesthetic emotion and a concept, the sublime involves the aesthetic emotion and behavior. MFT provides a foundation for the mathematical description of these abilities: similarity performs both of these functions; it establishes relationships among data and models (concepts of Understanding), and it activates actions of adaptation toward improving the harmony between the models and nature.

At this point, a reader might exclaim: you dragged me through four paragraphs of high-flying notions, but surely, your MFT examples in Chapters 5, 6, and 7 do not explain what is beautiful, there is no equation that can tell the difference between Rembrandt, Warhol, and a causal recreational artist. Of course not. The purpose here is to demonstrate that there is a possibility for "cyberaesthetics": a mathematical theory of mind in which concepts of emotion and beauty have a place. But then, why not proceed directly to producing by means of MFT even a simple example of something beautiful? Where is the difficulty that precludes this? The difficulty is in the adaptive nature of beauty. Kant got himself in trouble with later readers and admirers, by his attempts to provide examples of what is beautiful and what is not. His examples (such as, e.g., that drawings could attain pure beauty and paintings could not) were immediately criticized and he was branded as having undeveloped aesthetic taste and worse. One of the reasons is that what was beautiful thousands years ago is not necessarily beautiful today. Concepts of Understanding evolve, and those concepts that were useful some time ago, in that they captured important aspects of nature and provided an evolutional advantage to those who possess them, are not necessarily useful any longer. Within our evolving internal models, some concepts may become commonplace, outdated, empty of useful contents, and contrary to newer, better adapted concepts. Because mathematically, beauty is related to the harmony between the internal model and nature, it is changing with time. What is an excellent harmony between an adaptive model and data in an engineering system considered in Chapter 7 is a very simple construct, unworthy of the word "beauty" in the context of our mind. To design android-robots capable of human-level perception of beauty, even at a rudimentary level, their internal models have to be much more complicated than examples in this book. And, possibly, they would have to learn for many years, as humans do, to acquire an individual's subjective experience, and then their perception of beauty will acquire human-like, individual features. Section 10.3 considers directions for developing capabilities for the learning of complex internal models.

10.2.3 Instincts, "Lower Emotions," and Psychological Types

The mathematics of instincts and lower emotions is essentially the same as for the "higher" intellectual processes. And the reason is simple: the latter capability evolved on top of the former. Instincts and lower emotions operate with simpler models than thinking, but the

basic mechanism is the same. The Judgment faculty, first, must be able to identify sensory signals corresponding to desired objects, therefore, a similarity measure is a basis for the Judgment. Second, "lower" Judgment is fused with instinctual values. Recognition of food, sex, or danger objects leads to affect states characterized by the fused operation of the appropriate control systems and Kant–MFT system in a vortex of reciprocal connections, leading to a fusion of similarity with other instinctual values.

Fusion of "lower" level functions (instincts) and "higher" level functions (pure knowledge) is of a basic origin, and therefore persists in our psyche. Often, it is quite difficult to separate "pure" thinking from "lower" emotions. Varying degrees of fusion of various functions is responsible for varying individual characters or psychological types. According to Jung, there are several psychological types of mind, including a predominantly thinking type and another predominantly emotional type. Could these types of mind be explained by varying degrees of fusion among lower (instinctual) and higher (knowledge-related) judgment functions within MFT? Such an explanation might often be implied in a condecsending attitude of a scientist to an emotional-type person. Yet, many a scientist might have recognized a peculiarly rational and often superior adaptation potential of the emotional intellect. I think the above explanation does not fully account for the complexity of emotional intelect. A subtle variation of the above mechanisms is needed to account for the highly adaptive and rational nature of emotions in individuals of emotional psychological types.

The Jungian thinking-type mind has an a priori model oriented more toward nature. This "orientation" of the internal model corresponds to the original Kantian conception of the Understanding faculty. The Jungian emotional-type mind has an a priori model oriented more toward relationships with people. An internal model with this orientation, possibly, emphasizes certain a priori models of relating to other people and models of morals, which Kant conceived as a part of Reason. Models of each type can be of introverted or extroverted types. So "nature" includes external nature (physics) and internal nature (psychology and physics of mind). "Other"-people include oneself as well. (This is related to an ability to "see oneself from the outside.")

Emotional intellect if often oriented to relations among humans; it requires internal representations of another human being. At least, two semiindependent models or "archetypes" are needed, Self and Other. The Other serves for the internal projection of Self (emotional introverted experience), or for introjection of other human beings (emotional extroverted experience). Compared to thinking-type models (concepts of natural objects), the emotional-type models involve concepts of human relationships that are more directly related to instinctual emotions. Emotional intellect is related to our ability to endow certain concepts with high values, so that these concepts "act" as sources of emotional signals, similar to instinctual drives. These "emotional concepts" assume the role of differentiated adaptive instincts; they affect recognition and learning of other concepts. And they affect these thinking processes through the faculty of Judgment, which becomes integrated with "emotional concepts." Thus, "high emotional intellect" is based on emotions that are detached from the basic instincts and attached to certain concepts. These emotions, therefore, acquire highly differentiated and adaptive status, and are subject to will and adaptation. Yet, they are capable of immediately affecting judgment, like "lower" emotions related to basic instincts.

Modeling the complexity of the human psyche would require considering a complicated set of intricately interrelated models and emotions corresponding to various archetypes, and

modeling them on various levels of consciousness (from unconscious collective archetypes to individual unconscious, to threshold of awareness, to differentiated consciousness). This is a task for a future project. It could be compared in complexity to the human genome project; and if the current human genome project is to embark on mapping genes determining psychic human makeup, it would inevitably lead to such a task.

Future mathematical descriptions of complicated emotions will have to account for the lessons of the past. The main lesson is that the nature of emotions is fundamentally different from that of concepts. Attempts to describe emotions mathematically in a way similar to concepts lead to combinatorial explosion. The role of emotional signals in neural networks is to carry instinctual needs to the modules that control recognition and action. Recognition and understanding are not "disinterested," but are influenced by instinctual needs. "Higher" emotional intellect is related to high-value concepts that are capable of generating emotonal signals, and, like instinct, can affect Judgment. The emotional mechansim is implemented in MFT through the measures of similarity. Thus, it seems that complicated, structured measures of similarity, involving emotional models, will have to be developed to describe emotional intellect.

10.3 Learning: Genetic Algorithms, MFT, and Semiosis

10.3.1 The Origin of A Priori Models

Throughout the book, we have argued that learning is based on a priori internal models. But where the models came from? Do they include everything learned so far, or only what is learned in early childhood? Are they genetically inherited? Should we also consider genetic evolution as part of the learning process? The answer to all these questions seems to be yes, depending on the level of analysis. Classical philosophers from Aristotle to Kant considered "a priori" as prior to and transcendent to *any* experience. Today, being aware of genetic evolution, early childhood development, and the adaptive-evolutionary character of the very notion of truth, we are more careful about the definition of "a priori." Given the complexity of the problem, we should consider separately the nature of the a priori models and adaptation mechanisms at various time scales: individual neural firings, individual brain development, and genetic evolution. A current state of the internal model determines the behavior and the adaptation of a system over the next few moments, and, thus, plays the role of an a priori model at this time scale. The development of an individual mind is based on genetic information, which is an a priori model in this process. And genetic evolution proceeds from simpler life forms, or possibly from before life began and any a priori models were available.

Some researchers would argue that a priori models of living beings evolved from similar mechanisms found in inorganic matter: when two molecules interact and form a chemical bond, they "can do" it because each of them has a "model" of the other in the shape and structure of its electron wavefunctions. It might even make sense to say that the two molecules forming a bond "recognize" each other. This leads to a more profound discussion relating evolution and natural selection to unknown yet fundamental laws of physics. Internal models of living beings evolved in the process of evolution, guided by the law of natural selection. Should we conclude that in the evolution of our universe from the

Big Bang, the fundamental laws of physics, as we know them, are evolving guided by a yet unknown physical law? And does this more fundamental law of the general evolution of nature somehow "favor" selection of more complicated forms of matter? This discussion can also go in the other direction: the laws of evolution of living matter might be more intimately connected to the laws of physics than currently appreciated. Newton, were he alive, would love to ponder this!

In engineering applications we often start with fairly sophisticated models, trying to incorporate into the a priori model an expert-level knowledge, while providing for an adaptivity to unknown or unpredictable variabilities. Engineered intelligent systems with adaptive models attempt a mathematical description of intelligence operating with complicated a priori concepts and having a limited degree of adaptability. For example, prediction and tracking models described in Chapter 7 start with a few types of a priori models; they can learn to identify and track hundreds of objects in variable noise and clutter conditions; still, their adaptivity is limited to tracking specific types of objects.

Adaptivity of "engineering" models is limited: on their own, they cannot learn something entirely different from what they are designed for, such as riding bicycles or cooking meals. It would be quite desirable to be able to build "android" systems that can learn, like ourselves, various types of domain knowledge and appropriate behavior. Keep in mind that reaching an expert level of performance requires many years of schooling and experience in interaction with human experts. But even allowing for a long learning period, we do not know yet how to build mathematical systems with these "android" capabilities. And the big unknown is: what content of our a priori models enables this type of general learning? An ability to learn a human-type language seems to be crucial, because we think, to a significant extent, in terms of concepts contained in our languages (words, sentences, etc.). These concepts combine the necessary degree of specificity and fuzziness, and provide for both apriority and adaptivity. Beginning in the 1950s, a massive effort has been devoted in linguistic research to identifying the a priori models of language (linguistic faculty of mind). So far, this goal has not been achieved. A specific difficulty was in combining apriority and adaptivity; therefore, it is possible that MFT might facilitate the progress in linguistics.

Toward the development of "android" systems, remaining challenges include learning or "evolving" complex hierarchical structures from simpler ones. Previous chapters have considered the first steps toward this goal, in particular, identifying the number of submodels and combining submodels of several types. The next three subsections explore possible future directions of combining MFT, CAS, and semiotics.

10.3.2 Genetic Algorithms of Structural Evolution

Holland (1995) explores an explanation for evolution of complex models from simple ones based on the concept of genetic algorithms. In Chapter 2 (Section 2.13) we reviewed his concept of intelligent agents and their organization into complex adaptive systems (CAS) as well as genetic-type algorithms for learning and evolution. Here we briefly recall the main definitions, then discuss relationships between MFT and CAS, and directions for further development. In CAS, each agent is a "simple" if–then rule: if (**a**) then (**b**). The if-part of an agent tests conditions, and the then-part performs actions. A large number of agents are sending and receiving messages: some messages might come from sensors or can go

to actuators, but most devote their activity to an internal thinking process: building and estimating internal models.

Adaptation is achieved by (1) generating new rules and (2) selecting good rules and their combinations and discarding bad rules and their combinations. Genetic algorithms are used for rule generation; two types of algorithms have being considered for rule selection. New rules are generated by two types of genetic operators, crossover and mutations. Crossover acts in the process of "mating" of two parental agents: with certain probability, two agents mate and produce an offspring. An offspring is a new agent with a-message-receiver or b-message-transmitter obtained from the parental ones by a crossover operation, an exchange of substrings between two message-strings. Mutations act by a random replacing of a single character by a different one. Mutations are needed to retain adaptivity even within those substrings that came to dominate the population genome.

A first algorithm for the rule *selection* is a credit assignment algorithm, which is a variant of Adam Smith's capitalistic "invisible hand." According to this algorithm, agents within a CAS system are in competition for posting their output b-messages on a web page. They "bid" for a limited number of available slots and higher bids win. They have to pay with available "cash." Similarly, agents are in competition for using input information that they have to "buy" from the web page. Cash paid is credited to the posting agent. A second algorithm for rule selection is a genetic selection algorithm. According to this algorithm, the probability of mating among agents is proportional to their fitness. Fitness can be determined by direct "survival" in an environment, or by the amount of "cash" accumulated by each agent according to the first credit-assignment algorithm. Thus, offspring in each generation are expected to outperform the average fitness of the population.

In CAS systems and genetic algorithms, the unit of adaptation is not an individual agent, but a population or system of agents. The evolutionary "pressure" leads to the selection of "good" building blocks or schemata. Schema is a mathematical notion corresponding to a generalized concept of a collection of building blocks (or substrings) that coevolves in the process of evolution. A schema may include "ignored" positions (*); for example, schema 1#**#***** includes all strings beginning with 1# and having # in the fifth position. Schemata are not used in the algorithms, but for a mathematical analysis of genetic algorithms.

10.3.3 MFT, CAS, and Evolution of Complex Structures

Here we discuss relationships between CAS and MFT and outline future directions toward the development of complicated heterohierarchical intelligent systems. Let us compare the MFT partial conditional similarity measure $\mathrm{ll}(\mathbf{X}_n|\mathbf{M}_k)$ that determines a similarity between an nth piece of data and kth model, to CAS a-string defining a condition for the kth agent. A CAS agent is activated if data string $\mathbf{X}_n$ matches $\mathbf{a}_k$, so that in non-#-positions, $\mathbf{X}_n = \mathbf{a}_k$. Note that this is a simplified nonadaptive version of the MFT similarity. For example, consider Bayesian MFT with Gaussian pdfs, with the means, $\mathbf{M}_k = \mathbf{a}_k$, and diagonal covariance matrixes, $\mathbf{C}_k = \mathrm{diag}(c_{k1} \ldots c_{kD})$; then, $\mathrm{ll}(\mathbf{X}_n|\mathbf{M}_k) = -0.5 \; \Sigma_i \; (X_{ni} - M_{ki})^2/c_{ki} + \mathrm{const}(k)$. If we define c_{ki} to be very large for positions i for which $a_{ki} = \#$, and very small for all other positions, the MFT similarity $\mathrm{ll}(\mathbf{X}_n|\mathbf{M}_k)$ acts exactly like the CAS a-condition. Let us also define the MFT model parameters $\mathbf{S}_k$ to match the CAS b-message for the kth agent, $\mathbf{S}_k = \mathbf{b}_k$, and MFT model $\mathbf{M}_k$ to be nonadaptive, nondependent on parameters $\mathbf{b}_k$. Then

the effect of data $\mathbf{X}_n$ is the same in MFT and CAS systems: the condition (class, concept, model) k is activated and the message $\mathbf{b}_k$ is generated. (To emulate #-characters in an MFT-transmitted b-message, the transmitted message can also include b-covariance, defined in a manner similar to the covariance above, with large values in #-places.)

The above discussion illustrates that MFT agents can perform all the functions of CAS agents. On the one hand, the MFT agents, in general, are more adaptive and more powerful, are capable of adaptation at an individual level, and have a capability of aggregating (associating, segmenting) the input data-messages according to geometric, dynamic, and other type models. On the other hand, CAS possesses evolutionary capabilities and CAS agents have aggregational capabilities that can be used to build hierarchical models. The previous analysis shows that all the additional capabilities of CAS and genetic algorithms are applicable to MFT. So the two powerful techniques of parametric and structural adaptation can be naturally combined together. In particular, the evolutionary development of complex a priori models by structural aggregations of agents is being studied extensively for CAS systems.

Lets us compare in more details the adaptation and role of fuzzy logic in MFT and CAS. In Chapter 2 we analyzed the conundrum of combinatorial complexity vs. CAS systems and came to a conclusion that a unit of adaptation is not an individual agent, but a schema. We saw that combinatorial explosion is avoided in CAS by means of fuzzy logic acting at the level of schemata. Recall that combinatorial explosion is avoided in MFT by using fuzzy logic and fuzzy internal models. The evolution of schemata is more similar to MFT adaptation than learning at an individual agent level. An information representation in an agent is characterized by its "genetic code," which we denote $\mathbf{g} = (\mathbf{a}, \mathbf{b}) = (g_1, \ldots, g_L)$. (We consider each g_i as taking one of two values, 0 or 1.) Every g-code defines a crisp logical if–then statement. Consider now a proportion of each allele (a g_i value) in the population. It is given by an average value of g_i in the population, $r_i = \bar{g}_i$, and the average g-code is given by $\mathbf{r} = (r_1, \ldots, r_L)$. The population-average g-code, $\mathbf{r}$, is a fuzzy and not a crisp statement about the allele values. The uncertainty or fuzziness associated with each r_i is characterized by its variance, $c_i = r_i(1-r_i)$. If the entire population has $g_i = 1$, then $r_i = 1$, and the variance is zero (of course, the same is true if all $g_i = 0$). Thus, schemata can be viewed as fuzzy submodels obtained by averaging nonfuzzy individual-agent models over a population.

Similarly to MFT, this fuzziness has a dual role in CAS systems; first, it impacts an uncertainty of the system response to any message $\mathbf{a}$ with 0 or 1 in the ith position. And second, only fuzzy parameters can adapt. Positions i with nonzero variance has a chance to vary in the normal reproductive process (selection of the best fitted agents + crossover), if parents have different alleles in the ith position. Positions i with the variance of zero do not vary in the normal reproductive process and do not adapt, except by mutations. Recall that similarly in MFT, a zero (or very small) variance C_{ki} for feature i leads to zero association between a concept-model k and data that do not match the k-model perfectly (for i-feature). First, such a k-agent cannot be activated by nonperfectly matching data and second, such an agent loses its adaptivity, because it cannot react to new data (see Problem 10.3–1). So, as in CAS and MFT, the variance controls both the uncertainty about the data and fuzziness about the model parameter uncertainty (and their adaptivity).

There are more detailed similarities among CAS and MFT mechanisms. In MFT, a model is prevented from concentrating on a single piece of data and from loss of adaptivity

by preventing covariance from going to zero. CAS has two mechanisms to deal with these two problems. First, a model is prevented from concentrating on a single piece of data because the data and models are discrete (continuous variables are discretized to 0 or 1, so their uncertainty is not smaller than a discretization interval of 0.5). And second, adaptivity is maintained by mutations. Both systems allow for a stable learning of "good" models: mutations in CAS are rare, and an agent's variance in MFT may attain a small value, making further adaptation of this agent unlikely.

We would like to emphasize the relationship between fuzziness and the noncombinatorial nature of CAS adaptivity. CAS agents are nonfuzzy and nonadaptive. CAS schemata are fuzzy and adaptive. The genetic mechanism of preferential reproduction for better fitted agents creates a gradient in the space of parameters of fuzzy schemata leading to schemata adaptation. This gradient is in the direction of increased fitness, or increased cash (when a credit assignment algorithm is used). Fitness or cash can be thought of as allocated to schemata, by taking the average fitness (or cash) of agents belonging to each schema. There is a certain similarity between the cash–credit assigned to schemata and MFT similarity measure: cash is credited according to utility, and utility requires a match between the data and the model [data are X, or b-messages and the model is M, or a-message; similarity in CAS is of the form if$(X = M)$ and in MFT, ll$(X|M)$]. Therefore, similarity governs credit assignment and is playing the role of "universal currency" among agents. Thus we came to a concept of a parametric adaptive evolving model of structure.

Comparing what is known today about the brain and chromosomes, it seems, on the one hand, that MFT is a more plausible mechanism of mind than genetic algorithms. It will be interesting to obtain direct psychoneurological evidence for each type of adaptation mechanism of mind: for genetic-type recombination learning and for MFT-type gradient-and-association learning. On the other hand, in the area of genetics proper, genetic algorithms seem to be more plausible. There is no evidence for feedback from phenotype to genotype at the level of an individual organism (i.e., genetic inheritance of acquired features), which is necessary for gradient learning. In fact, it is a sacrilege to raise such an issue among geneticists. However, gradient learning provides such a huge advantage in adaptivity that it is difficult to believe that nature does not use this mechanism at all in genetic evolution. (And how much, often in vain, would we like our children to inherit what we learned so hard.) It is clear that such mechanisms, if they exist, should be very (very!) subtle and could act on our genes only with extreme caution. Otherwise, too much feedback would quickly modify the genetic information accumulated over billions of years and would lead to overspecialization to a particular environment. Then a sudden change in that environment (physical, chemical, or bacteriological) could lead to extinction. Did not something of this sort happen during the Cretaceous period leading to an extremely fast adaptivity and proliferation of a tremendous number of species of dinosaurs and other creatures, most of which could not adapt to a sudden change in environment? Could it be that the adaptation mechanisms of survival became more cautious since? Still, subtle feedback mechanisms cannot be ruled out and geneticists are waking up to such a possibility. One indirect mechanism of this type, gene mutators, was discovered recently: there are specific mutator genes that can tremendously speed up adaptation by increasing mutation rates of certain other genes. Nonrandom mutations directed toward better adaptation to the environment have been observed over one generation. The genetic adaptation, therefore, can occur without selection mechanisms acting at the phenotype level. The mechanism

for such adaptation can be described as "genetic cell selection." If this can occur due to environmental influences, the possibility that genetic modifications can occur under the influence of the organism cannot be excluded.

There are many unknowns in the actual operations of genetic mechanisms in nature. For example, a large part of DNA is inactive; it does not participate in protein synthesis and it does not affect the phenotype. Mutations that accumulate in this part of DNA do not affect the fitness of organisms, and it seems to be free from normal evolutionary pressures. Could this be a laboratory in which the models of our future internal models are being developed with some subtle feedback mechanisms? We will be better able to answer these complicated questions about adaptivity of internal models in the genes and in the mind once we better understand the mathematics of combining structural and parametric dependencies: *parametric structures.* Some of the directions for such future research are outlined in this section.

The above comparison of CAS and MFT accentuates a philosophical difference between the two systems. CAS systems and genetic algorithms are not "nice": individual agents do not learn, they die with the same internal models as they are born with, only the population learns and evolves. The nature of learning is antipersonal with regard to individual agents. MFT agents are much more "personalistic," there is a personal learning at the level of individual agents. This learning is possible due to a concept of similarity, which Kant called Judgment and considered as a foundation for higher intellectual abilities, including higher emotions and beauty. Beauty, according to Kant, is an ability to perceive purposiveness (of our internal representations in their relationships to the outside world) as divorced from a specific fitness-type goal. In MFT, similarity is an ability of this type. It would be very interesting to demonstrate the evolution of simple CAS agents toward the concept of similarity (say along the lines of our previous discussion), and then toward the concept of beauty.

10.3.4 Semiosis: Dynamic Symbol

The popularity and subsequent fall out of favor of "Symbolic AI" left many researchers antagonistic toward the word "symbol." But an impression that "Symbolic AI" represented the mathematics of symbols is very wrong. A *symbol* is not a monumental piece of bronze sitting on a foundation of stone. A symbol is a fleeting vortex of an interacting perception, feeling, a priori model, adaptation, attention, behavior, and concept formation. In other words, a symbol is a process. It is a process of thought.

At the beginning of this chapter (in Section 10.1.1) we considered an example of a lobster finding food in the ocean. Let us repeat: there is no such thing as "food" in the ocean, "food" is a dynamic process of interaction between an object in the ocean, sensing neuron (that forms an internal representation-signal), grabbing neuron, etc. For a lobster, "food" is a symbol including the whole process of sensing, recognizing, grabbing, and eating. A complex psychological notion of symbol was discussed by Carl Jung; the Jungian symbol is a thought process connecting consciousness with unconscious archetypes.

Semiotics is a science devoted to studying signs and symbols that was overviewed in Section 2.12. Let us briefly summarize. The founders of semiotics, Peirce and Morris,[4] introduced the notion of a sign as a trilateral unity of sign-vehicle (the media used as a sign), designatum (the object that the sign refers to), and interpretant (an internal representation of

the interpreted or recognized sign). The intelligent system or mind interpreting the sign is called an interpreter. We use a slightly different terminology: *sign* is used for sign-vehicle, and *symbol* is used for the trilateral process of sign interpretation. Morris decomposed this process into the three dyadic relationships-processes: syntactics (relations among signs), semantics (relations between signs and their designata), and pragmatics (relations between signs and their interpreters). The symbol-process of interaction within the triadic unity is a part of *semiosis*, the learning (adaptation, evolutionary) process at a system level involving multiple triadic symbol-processes.

We use *sign* for nonadaptive entities and *symbol* for an adaptive process of sign interpretation. This corresponds to Pribram's designation of "signs" within the brain as acts of communication that are invariant to the context, while "symbols" are context dependent. He understands signs as the results of the associative cortex affecting the input sensory systems; that is, a process of perception. These are the a priori, less-adaptive aspects of the internal models, which nevertheless are affected by past experience, thus having a degree of adaptivity. Symbols, according to Pribram, are the results of interaction between the frontal lobes (concepts) and limbic system (emotions); they are stimulants to actions and are sensitive to context. Frontal lobes are responsible for conceptual understanding and the limbic system is involved in emotions. Therefore, Pribram's analysis of symbol processing in the brain corresponds to our definition of symbols as processes involving emotions and concept formation. Our definitions, corresponding to the mathematical apparatus of MFT, emphasize the common nature of perception and cognition, and more consistently differentiate signs and symbols. Our definition emphasizes an important role that is allocated to symbols by the general culture and by psychology: symbols are creative processes, bringing into consciousness the unconscious fuzzy a priori models. Symbols expand the conscious aspect of the internal models.

MFT agents give the mathematical description of the dynamic symbol-formation process. The symbol process is equivalent to a loop of the Kantian triadic-mind process of Understanding–Judgment–Reason. In relating our analysis to psychology, we call these loops "vortexes," vortexes of thought. The relationships between the semiotic description of the symbol-semiosis and mathematical description of the Kant–MFT agent-loop are illustrated in Fig. 10.3-1 and Table 10.3-1. A sign is a subset of input data, a structure in the data. It refers to a designatum, an object in the world (or a signal from another agent). An interpretant is an internal representation, an output signal from the Judgment module indicating that a unity (high degree of similarity) was established among a sign, object, and model. An interpretant is sent to the Will–Reason module within the same agent, where it is used to sustain the loop of adaptation of the model to the sign and object. An interpretant is also sent to other agents as an input data. The interaction with other agents establishes the meaning of the interpretant and its object-designatum. The interpreter is the entire intelligent system, including multiple agents.

In interaction with Reason and, even more important, with other agents, an interpretant acquires meaning; thus, both Judgment and Reason are within the domain of semantics. Reason produces actions of two types: within the agent and outside of it. Within the agent, Reason modifies the internal model, and adapts the model to improve its correspondence to the sign. Outside the agent, Reason acts on the world (or other agents), possibly leading to changes in the data and in signs. Reason and the actions it generates are the domain of pragmatics. Syntactics refers to relationships among signs. Interpretant-output from a

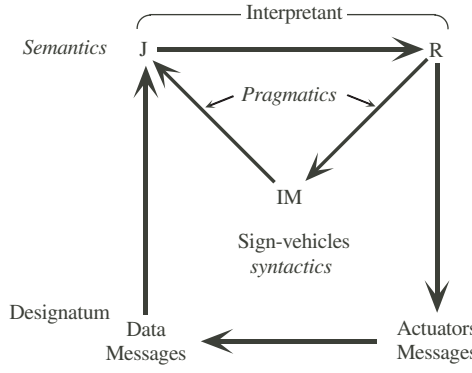


Figure 10.3-1 Relationships between the semiotical description of the symbol-semiosis and the mathematical description of the vortex of symbol formation in the Kant–MFT agent-loop.

particular agent serves as input data to other agents. From the point of view of these other agents, interpretants are signs, and there are syntactic relationships among them.

There is a correspondence between internal models and signs in the input data. Therefore, structures and relationships among the internal models to some extent parallel those of signs and syntactic relationships among them.[5] Relationships among individual-agent models are what Kant called Understanding. Models include a priori, fuzzy models with low similarity to the data, and crisp models, which have already been adapted to the data and whose correspondence to specific data-structures has been established with a high measure of similarity. Crisp models are the concepts of Aristotelian logic. Fuzzy models are fuzzy concepts, and a large number of models with varying degrees of fuzziness are always present in the mind. A priori, highly fuzzy models of primordial origin are the Jungian archetypes. They are not directly accessible to consciousness; conscious contents include mostly crisp and low-fuzzy models.

TABLE 10.3-1
Relationships among Semiotics, MFT, and Kantian Theory of Mind

Semiotics	MFT	Kantian Theory of Mind
Sign-vehicle (sign)	Subset-structure in input data	(Kant did not analyze structures within the world-manifold)
Designatum	Associated model and data subset	Phenomenon
Interpretant	The output signal indicating a recognized concept (high similarity measure)	Judgment result
Syntactics	Structures/relationships in the data	(Kant did not analyze structures within the world-manifold)
	Relationships among models	Understanding
Semantics	Relationships among agents, similarity measures, and behavior generation (including adaptation laws)	Judgment + Reason
Pragmatics	Behavior generation (including model adaptation) and dependence of similarity on models	Reason
Interpreter	Intelligent system	Mind

NOTES

1. Some researchers think that the antinomy is resolved by "talking it away." They consider gradually increasing levels of complexity of mind, behavior, and adaptivity, and think that in this way they can talk about gradually increasing levels of freedom. But this is a misunderstanding. A door on hinges can be considered more "free" than a concrete wall, but this has nothing to do with the concept of freedom as a fundamental concept of our existence. Causal explanation at every level eliminates the possibility of freedom, as perceived by every human being. Kant assigned *causation* to the domain of Understanding and *freedom* to the domain of Reason, so that one does not have to be related to the other. The problem is in actually explaining what happens at a high level of complexity, so that our notion of freedom can be reconciled with a "complex" causal explanation. Currently this does not seem possible.
2. Neural and cognitive sciences have been concerned with relating emotions to material neural and bodily physiological functions. For example, neural pathways have been found from the hypothalamus (brain areas associated with emotions) to viscera. The popular known connection between fear and an upset stomach could possibly be understood as a survival mechanism regulating interactions between fear and hunger (this connection has been observed in primitive animals as well, e.g., such fierce predators as Commodo Dragon lizards). This is an example of the "lower" aspect of emotions. Brain research relating emotions to higher intellectual functions is yet in the incipient stage. Interactions between cortical systems (associated with high cognitive functions) and the hypothalamus are hypothesized to be mediated through the amygdala. The high degree of reciprocal anatomical connections found among these neural structures in the human brain suggests the existence of information-processing loops involving emotional and cognitive functions (such as our MFT–Kantian loops). Our knowledge of the structure of the brain is insufficient for deducing the mathematical theory of high emotional functions, and many believe that even if the entire wiring diagram of the brain were available, it still would not be possible to deduce its main mathematical concepts. Neural and psychological data have to be combined with philosophical analysis and physical intuition to develop a mathematical theory of higher emotions.
3. This is why discovering harmony within the chaos, such as fractal theory, is especially pleasing.
4. Concepts of semiotics date to Ancient Greece, so calling Peirce and Morris the founders of semiotics may not be accurate.
5. Let us remember that an "internal model" has three aspects: the concept-model, the meaning of which is established by its structural relationships with other models, the image-model, which is similar to the data subset, and the process-model, which generates the image-model. In the present discussion we refer to the first aspect, the concept-model. The concept-model is closely related to the interpretant: the interpretant indicates that an object corresponding to the concept-model is recognized.

BIBLIOGRAPHICAL NOTES

Kant theory of mind (Kant, 1781, 1788, 1790).
Fuzzy logic (Zadeh, 1962, 1965).
Relationships between fuzziness and hierarchy (Meystel, 1995).
External representations in the world (Freeman, 1996).
Genetic algorithms and complex adaptive systems (Holland, 1992, 1995).
Genetic mechanisms of DNA (Frank-Kamenetsky, 1996).

Semiotics' foundation publications (Peirce, 1935–66; Morris, 1971).
Dynamic symbol as a vortex of concept, feeling, and action (Dmitriev and Perlovsky, 1996, 1997).

PROBLEMS

10.3–1 Show that in MFT, an agent that loses its nonfuzziness (variance goes to zero) loses its adaptivity. *Hints:* Consider fuzzy association definition (4.3-13) for Gaussian pdfs (4.3-6). Show that for $C_k \to 0$, $f(k|n) \to 0$ or δ_{kn} if there is $X_n = M_k$ (that is, this agent concentrates on a single piece of data n). Consider MFT adaptation Eq. (4.2-3). Show that if $f(k|n) = 0$, $d\mathbf{S}_k/dt = 0$; and if $M_k = X_n$, consequently $\partial\mathbf{M}_k/\partial\mathbf{S}_k = 0$, and $d\mathbf{S}_k/dt = 0$.

10.3–2 Relate discussions in Section 10.3 to the intelligent tracker discussed in Section 1.1.4.

This page intentionally left blank

part three

FUTURISTIC DIRECTIONS

Fun Stuff: Mind—Physics + Mathematics + Conjectures

This last part of the book explores future research directions toward a physical theory of mind. Is there a fundamental difference between human and machine? What is consciousness? Is a physical theory of consciousness possible? What are the limits? Is physics of mind related to yet undiscovered mysteries of material substance? Chapter 11 considers Gödel's theory and Penrose's arguments and discusses their relevance to a physical theory of mind. Chapter 12 discusses consciousness and the possibility of a physical theory of consciousness based on modeling field theory. I attempt to delineate the current limits of the scientific method. The Epilogue presents still a fresh view on the contents of the book and future directions.

This page intentionally left blank

chapter 11

GÖDEL THEOREMS, MIND, AND MACHINE

Gödel proved that formal systems related to Aristotelian logic or logic of predicates are fundamentally limited. Turing has reformulated this result for computational systems. There have been several attempts to use these results to prove the principled difference between the mind and machine. A recent one by Penrose, who believes that the Gödel–Turing limitations have to be surpassed to model the mind. But Penrose's mathematical arguments are considered fallacious by many mathematicians. Is this discussion relevant to the philosophy of mind? Is it relevant to building practical intelligent systems? This chapter briefly reviews Gödel–Turing results, Penrose's arguments, and some counterarguments. We analyze the combinatorial nature of Aristotelian logic as revealed by the arguments of Gödel and Turing and compare it with the combinatorial explosion of complexity of intelligent algorithms and neural networks. Our conclusion is that Gödel–Turing results establish limitations to Aristotelian logic, but are not necessarily relevant to the theory of mind.

11.1 Penrose and Computability of Mathematical Understanding

Penrose came to the conclusion that conscious understanding cannot be explained as a feature of a computational system. Consciousness could not have arisen as a result of computations, nor can computation ever simulate consciousness. Although considering mathematical understanding, he believes that any understanding and consciousness in general cannot be explained as a computational activity. The brain activity responsible for conscious understanding, according to Penrose, "must depend upon a physics that lies beyond computational simulation."

Wigner once delivered a lecture entitled "The unreasonable effectiveness of mathematics in the physical sciences." This theme represents one side of Penrose's arguments. Why should the physical world be describable by an abstract mathematical construct? And why could such an abstract mathematical construct be knowable to the human mind? Let us take a view that our mind has evolved from unorganized matter in the process of genetic evolution, for the purpose of controlling our body and our behavior, guiding it toward improved survival. This point of view explains a tremendous number of facts

in paleontology, anthropology, and psychology. It explains, in particular, why our mind evolved toward physical understanding of the world: because an ability to predict events and their interrelationships is beneficial for survival. But if we take this view consistently, we have to acknowledge that an approximate understanding, say at the level of one-tenth or one-hundredth, of what actually occurs should suffice for most survival-related activities, such as hunting, home building, etc. Why should our mind be able to generate physical theories accurate to one in a millionth of a millionth part? This seems impossible to explain by an evolution postulate. Nor does it seem possible to explain why the physical world ought to follow precise mathematical relationships. Penrose feels compelled to accept a modified Platonic view in which there are three interrelated worlds: the world of consciousness, the world of matter, and the world of ideas, including mathematical objects and constructs.

The other side of Penrose's arguments is based on the mysteries of the physical world of matter. Even though quantum theory and the general theory of relativity are capable of explaining and predicting physical phenomena with tremendous accuracy, there are unexplained and inconsistent aspects of these theories. In particular, the nature of the process of quantum measurement remains unexplained despite the concerted efforts of several generations of physicists. Quantum measurement refers to a process of relating a state of a quantum system to the macroscopic classical world accessible to our conscious perception. A quantum system is described by a wavefunction, which is a superposition of multiple states. For example, an electron traveling from one place to another can be described as a superposition of multiple trajectories. During a "macroscopic observation," these multiple states "collapse" to a single macroscopic classical state. Existing quantum theory provides us with a mathematical technique leading to an extremely accurate description of the observed physical world. But conceptual difficulties remain. One such conceptual difficulty was discussed by Schrödinger and goes under the name of Schrödinger's cat. Say that one of many electronic trajectories goes near a loaded gun and triggers a shot that kills a cat. Then, before a measurement, the system will be in a superposition of a dead cat and alive cat. Physicists do not believe that such states are possible. Somewhere on an intermediate scale between the quantum microworld and observable macroworld, wavefunctions should collapse, but nobody knows yet how to describe this mathematically.

Combined effects of quantum and relativistic theories could get even more strange, leading to the possibility of back-and-forth time travel. In the classical theory of relativity, properties of space–time, such as its curvature, vary depending on its physical state. Since a quantum system normally is in a superposition of multiple states, the space–time also is in a superposition of multiple states. In particular, there is a nonzero probability of a space–time state that curves so much that a closed time-like line appears. A closed time-like line means a possibility of traveling back and forth in time. Even if this effect occurs only at microscopic distances, it might still influence macroscopic systems. If closed time-like lines could be exploited in a computer, there is the possibility that the computer has at its disposal the results of computations before the computation began. According to Penrose this opens the possibility for noncomputable physics.

Penrose believes that new undiscovered yet physical principles of the material world are needed for the description of consciousness. A discovery of these new principles will constitute a theory that he calls Correct Quantum Gravitation (Penrose, 1989). This future theory will unify quantum theory and the general theory of relativity and will explain the nature of quantum measurement as a nonlocal, nonalgorithmic process. The nonalgorithmic

nature of the future physics, according to Penrose, will resolve the mysteries of creativity and free will, related to the exit out of the finite world of events into the infinite world of ideas.

11.2 Logic and Mind

One of the roots of contemporary approaches to modeling intelligence is logic. It used to be considered equivalent to intelligence. Logic is at the basis of most of algorithms of artificial intelligence. It dates to Aristotle, and its mathematical foundations were developed at the end of the nineteenth and the beginning of the twentieth century. This development revealed a fundamental limitation that was first seen by Cantor, then formulated in a famous Russell paradox, and finally was proved in several theorems by Gödel. Let us overview this development in which philosophy was interwoven with mathematics.

According to Kant, the a priori concepts of mind are purposive, in that they contain an intuition about the world. In the previous chapter, we saw that this purposiveness is related to adaptivity and requires fuzziness of the a priori concepts. This understanding did not exist in the nineteenth century and the mathematical formalization of the Kantian intuition met with difficulties. These difficulties were related to verification of the truth of mathematical statements concerning infinite objects. Hilbert thought to resolve these difficulties on a nominalistic basis. He developed an approach named formalism, which rejected the intuition as a matter of scientific investigation and formally defined scientific objects in terms of axioms or rules. Weyl argued that "A science can only determine its domain of investigation up to an isomorphic mapping. . . . The idea of isomorphism demarcates the self-evident insurmountable boundary of cognition." Isomorphism here refers to the multiple "real-world" objects that satisfy the same mathematical axioms and therefore are equivalent from the point of view of mathematical formalism. As summarized by Webb, formalism consists in that "being unable to intuit or know the objects of science in themselves, we must settle for the formal laws they satisfy." This is opposite to the position of Newton, who relied on his intuition about the world. On the one hand, formalization of mathematics divorced mathematics from physics, which relies essentially on physical intuition about the world of matter. But, on the other hand, it stimulated the development of sophisticated abstract mathematical methods that promised to solve forever the issue of mathematical truth.

Hilbert distinguished three stages in the development of any mathematical theory: (1) an informal theory, (2) a strictly formalized theory, and (3) a metatheory proving the consistency of (2). From the point of view of Kant–MFT theory stage 1 is the dynamic process of symbol formation, the process of adaptive differentiation of concepts; stage 2 is elaboration of Aristotelian logic in the transcendental domain of Understanding; but Hilbert's stage 3 is an attempt to prove the transcendental by using Aristotelian logic, which was later proved impossible by Gödel. In Kant–MFT theory, I replaced it with a new stage 3, a synthesis through continuous adaptation of differentiated concepts in an evolving heterohierarchical system. Whereas formalism wanted to save mathematical "infinite" arguments by formalizing them, an opposite approach to the foundations of mathematics, intuitivism, attempted to resolve mathematical difficulties involving infinite objects by ruling out infinities and requesting explicit definitions of mathematical objects

in terms of finite Aristotelian logic. In spite of prominent mathematicians contributing to intuitivist foundations of mathematics, it never quite succeeded. The limitation of the intuitivist approach seems to be related to its philosophical inconsistency: an attempt to utilize Aristotelian logic belonging to the domain of Understanding in the domain of Judgment that requires adaptivity and fuzziness.

In 1884, Frege introduced a procedure for defining numbers in terms of sets. First, consider all equivalent sets1 (whose elements can be one-to-one paired). Then, consider sets2 containing all equivalent sets1. An integer number n is a set2 of all sets1 containing n elements. This is the mathematical expression of formalistic isomorphism in which the number is formally defined in a most abstract way. In 1902 Russell exposed an inconsistency of this procedure by introducing a set R as follows:

$$R \text{ is a set of all sets that are not members of themselves} \qquad (11.2\text{-}1)$$

Is R a member of R? If it is not, then it should belong to R according to definition (11.2-1), but if R is a member of R, this contradicts the definition. Thus, either way we get a contradiction. This became known as Russell's paradox. For the next 25 years mathematicians where trying to develop a self-consistent mathematical logic, free from the paradoxes of the type noted above. But, in 1931, Gödel proved that it is not possible.

In 1937, Turing, developing a mathematical theory of computation, established a fundamental correspondence between computation and logic, formalized the definition of our understanding of algorithmic computation, and proved fundamental limitations of algorithmic computations similar to the Gödelian limitations on logic. These series of results by Gödel and Turing I will refer to as GT results, proofs, etc. In a simplified form these results are summarized in the next section.

Note that the idea of defining concepts (numbers in this case) in terms of sets is related to nominalistic philosophy, in which concepts are learned by noticing similarities among objects and defining concepts as sets of similar objects. In Chapter 2 we related this philosophy to specific computational concepts of intelligence and analyzed in detail the resulting combinatorial explosion of computational complexity. Throughout this book we have developed a different mathematical concept of intelligence, modeling field theory, based on the a priori origin of concepts and their adaptive relationship to the real world.

There have been several attempts to use GT results for establishing a fundamental difference between a computing machine and the human mind. A recent one is by Penrose, who presents "compelling arguments" based on GT results that mathematical and physical understanding could not be explained through computation. Penrose considers our consciousness or awareness of the properties of natural numbers, our ability to mentally visualize solutions to complex problems to be noncomputational, and this noncomputability to be demonstrably related to GT results. Penrose believes that the mysteries of the mind would be eventually understood through yet unknown physical theory, which will explain consciousness, creativity, and free will along with yet unexplained properties of the phenomenon of quantum measurement.

Penrose's arguments are considered fallacious by many mathematicians. I summarize Penrose's arguments and counterarguments by Putnam, which are related to the plausibility that an idealized mathematician's mind may not attain a thorough understanding of mathematical arguments in a perfectly conscious way.

11.3 Gödel, Turing, Penrose, and Putnam

Gödel's incompleteness theorems state that consistency cannot be proved within a formal mathematical system (powerful enough to include arithmetic).[1] "Formal" here means that the axioms and rules of inference are precisely described, so that they can be coded into a computer algorithm. This notion of coding mathematical rules into a computer was explicitly developed by Turing (1937).

Now I briefly summarize a Turing version of Gödel's theorem following Penrose. Consider a list of all algorithms $Cq, q = 1, \ldots$. An algorithm is a computer code (valid or not), which we assume can do the following: (1) given an input $n (n = 1, \ldots)$, an algorithm acts on it, that is performs a computation $Cq(n)$, and if this computation successfully *stops*, the algorithm writes out an answer, (2) otherwise an algorithm either never stops or fails to stop properly; in all these cases we say that the algorithm *does not stop*.

The above scheme is fairly general: it includes all possible algorithms, including learning or adaptive algorithms, and all possible input data. All algorithms can be ordered in a single list as above, because any algorithm can be coded as a finite computer code, and all such codes can be ordered, say in an alphanumeric order. This list is infinite, because algorithms of *any* finite length are allowed. Also, the above scheme accounts for algorithmic actions on any (finite) data, because any possible set of input data of any finite length can be enumerated by a single number, n.

Turing formulated a Gödel-type theorem in terms of a universal procedure A, that could decide if a particular computation, $Cq(n)$, never terminates. His conclusion was that there is no such procedure. The proof goes as follows. Assume there is such a procedure A that can act on $Cq(n)$, and A *stops* only when $Cq(n)$ *does not stop* (otherwise A does not *stop*). Since there is a list of all $Cq(n)$, A acting on $Cq(n)$ is an algorithm specified by two numbers (q, n), $A(q, n)$. Thus, the assumption is

$$\text{if } A(q, n) \text{ stops, then } Cq(n) \text{ does not stop} \tag{11.3-1}$$

Consider $A(n, n)$; this is an algorithm acting on a single number n, thus it is one of the original list of $Cq(n)$, for some $q = k$, $A(n, n) = Ck(n)$. Examine $n = k : A(k, k) = Ck(k)$. Substituting this into (11.3-1) we obtain

$$\text{if } Ck(k) \text{ stops, then } Ck(k) \text{ does not stop} \tag{11.3-2}$$

$$\text{if } A(k, k) \text{ stops, then } A(k, k) \text{ does not stop} \tag{11.3-3}$$

From (11.3-2) we conclude that $Ck(k)$ does not stop. But A fails to ascertain this because from (11.3-3) $A(k, k)$ does not stop.

Thus, we know that $Ck(k)$ does not stop, but there is no universal procedure A that would ascertain this. From this, Penrose concludes that human capacity for mathematical thinking is beyond any formal system or any algorithm whatsoever. Our capacity to conclude that $Ck(k)$ does not stop, according to Penrose, is due to a special mental representation, called mathematical understanding, which was shown to be of a noncomputational nature.

Let me briefly summarize the counterargument by Putnam (1995), who considers Penrose's argument fallacious. To complete Penrose's argument it is necessary to assume that

we *know* that A is a sound (valid) mathematical procedure (11.3-4)

Moreover, to compare an algorithm A to a human mathematician, we consider A encompassing the entire human knowledge relevant to this area of mathematics. There is no disagreement about this statement between Penrose and Putnam. But whereas Penrose believes that the nature of mathematical understanding is such that (11.3-4) is highly plausible, Putnam, insists that this is not necessarily so:

> A program which simulated the brain of an idealized mathematician . . . we might not be able to appreciate it in a perfectly conscious way, in the sense of understanding it or of being able to say whether it is plausible or implausible that it should output correct mathematical proofs and only correct mathematical proofs.

What is missing in Putnam's argument is an explanation of the nature of mathematical and physical intuition.

11.4 Gödel Theorem vs. Physics of Mind

Penrose develops a specification for a Gödelizing Turing machine, which is an explicit procedure that, given any algorithm A, constructs a computation C that we know does not terminate, but on which A fails. He shows that the number of binary digits in explicit specification of C is such that it exceeds this number for A by not more than

$$\Delta \sim 527 + 210 \ \log_2(N + 55) \tag{11.4-1}$$

where N is the number of internal states of the Turing machine implementing algorithm A.

Penrose's specification and expression (11.4-1) suggest the following interpretation. In a complete list of algorithms (Turing machines) needed for the formulation of the GT proof discussed in the previous section, there should be on the order of 2^{Δ} algorithms specified between the algorithm A and the computation C. Since $\Delta \sim 1700$, the explicit construction of GT proofs is possible in systems containing on the order of $2^{1700} \sim 10^{500}$ algorithms. This number exceeds all elementary particle interactions in the entire history of the universe ($\sim 10^{400}$). Thus it seems clear that GT limitations are not applicable to a human mind or to artificial intelligent systems.

The huge, inconceivable number of logical predicates that appears in the construction of GT proofs seems to be due to the inherently combinatorial nature of the constructs of formal systems, which consider a large number of possible combinations of predicates. It is interesting to compare this combinatorial explosion to a similar phenomenon in the design of algorithms of intelligent systems. In Chapter 2 we discussed in detail that the problem of seemingly inexorable combinatorial complexity plagued designers of intelligent algorithms for many years. We related this problem to the factors of adaptivity (an ability of an intelligent system to adapt to an ever-changing world) and apriority (an ability to utilize complicated a priori knowledge). And we traced it to a contradiction between Aristotelian logic and Aristotelian conception of mind (theory of Forms).

Consider the list of all algorithms Cq, that appears in GT theorems as containing an algorithmic representation of the internal model of an intelligent system, with various Cq being submodels, and consider each row, $Cq(n)$, as a result of actions or matches of internal submodels to the outside world, whose states are enumerated by n. This two-dimensional list is combinatorially long in both directions q, n. This can be viewed as a result of Aristotelian logic being inherently combinatorial, unsuitable for an efficient representation of the world, and lacking a capability for learning and efficient matching to the world.

In multilevel hierarchical intelligent systems, a procedure at a particular level in the hierarchy has a meaning at the next level. GT theory argues along the same direction, as follows. Our conclusion that $Ck(k)$ does not stop rests on our understanding of the meaning of how this procedure was constructed. When we encounter a particular $Ck(k)$ type statement (algorithm), let us explicitly include in the list of Cq the fact that it does not stop. This is similar to including the next level of the hierarchy. The new expanded (infinite) system of algorithms has its own Gödel-type $Ck'(k')$ statement, and our conclusion that it does not stop again comes from the understanding, which is outside of the systems. The procedure of expanding the system by including Gödel-type statements can be continued infinitely, leading to an ever-expanding hierarchical system.

Let us summarize the discussion of GT results as they are related to the combinatorial explosion of Aristotelian logic. GT reveals that the difficulties of Aristotelian logic are related to its excessive precision: any two concepts are either different or both are the exactly the same single concept. No matter how "small" the difference is between any two concepts, they are different. And an infinite number of concepts with even "smaller" differences can be "inserted" between them. I used quotation marks in these sentences, because there is no measure of difference within the standard Aristotelian logic. Fuzzy logic eliminates excessive precision inherent in Aristotelian logic.

Penrose expressed a deep-seated intuition that the human mind cannot be modeled by a computer in its entirety. This intuition is shared by many philosophers. But where is the border separating science from mystery, the forefront that challenges our rational understanding, physical intuition, and mathematical methods? We hope that with the development of the science of the mind, this border will continue moving. The discussion in the preceding chapters touched on a 2300-year controversy surrounding the issue of realism and nominalism, apriority and adaptivity of mind, which is being resolved by mathematical concepts combining an a priori adaptive internal model. Resolution of this long-standing controversy will affect our understanding of many phenomena considered until recently beyond computational modeling and physical intuition, such as meaning, qualia, the nature of understanding, and consciousness.

The nature of understanding in Penrose's sense of a specific awareness of the entire mathematical or physical theory ought to be analyzed as a differentiated set of phenomena, some aspects of which can be understood due to complicated properties of the internal representation or model, and other aspects of which are related to the expansion of the internal model and represent a challenge to our contemporary rational understanding. Further analysis requires that we consider consciousness as a differentiated set of phenomena in their interactions with unconscious. Within the concept of an internal model, we should consider the model of the outer world as well as the model of the intelligent system's self. On every level, the model of self contains a submodel of the previous level model-of-self,

resulting in a pyramid of self-reflections (Meystel, 1995). This pyramid of self-reflections as well as the pyramid of the outer-world model includes conscious and unconscious aspects at every level. What is the nature of creativity that expands the a priori model, not only its conscious content? What is free will, that strange feeling of purposeful indeterminacy that defeats any rational attempt at defining it? Would we find its explanation in new physical phenomena of a noncomputable nature postulated by Penrose? We continue this discussion in the next chapter.

NOTE

1. A conclusion is that within any such mathematical system there are undecidable statements, so that the statement (S) and its negation ($-S$) are both provable or true. It follows that any statement whatsoever ($S1$) is true. Here is the proof: if S is true then (S or $S1$) is true. If (S or $S1$) and ($-S$) are both true, it follows that $S1$ is true. This formulation emphasizes the dramatic consequences of Gödel's theorems.

BIBLIOGRAPHICAL NOTES

Penrose's views: on future physical theory of quantum gravitation (Penrose, 1989); on Gödel's theorem (1994, pp. 72–76, 117–126).

Russell paradox (Russell and Whitehead, 1908).

Historical discussion of the Gödel's theory (Webb, 1980).

Pyramid of self-reflections (Meystel, 1995).

Philosophical arguments why the human mind cannot be modeled by a computer in its entirety (Searle, 1980).

TOWARD PHYSICS OF CONSCIOUSNESS

What is consciousness? Why is it needed in biological or artificial systems? Can it be understood as a physical phenomenon? Can it be described mathematically? I outline a future modeling field theory of consciousness. In this theory, consciousness is due to an internal model. The specific internal model responsible for consciousness is what psychologists call Ego. The Chapter overviews the phenomenology of consciousness and Ego. It begins with popular conceptions and misconceptions and then continues the analysis, relating phenomenology to the modeling field theory. Even though our direct and naive perception of consciousness emphasizes its unity, wholeness, and predominance within psyche, consciousness is not a simple phenomenon. It is a complicated differentiated phenomenon, which cannot be described in isolation from the rest of the psyche. The phenomenology of consciousness is described in its intimate connection to the rest of the psyche, including the unconscious and emotions. Hypotheses and historical evidence concerning the origins and evolution of consciousness are summarized. Properties of consciousness are related to and explained within the modeling field theory. We overview neural structures involved in consciousness and emotions and identify candidate neural correlates for the modeling field theory modules and for Kantian theory of mind.

The discussion continues toward more complex aspects of consciousness, including the nature of creativity and free will. I analyze the differentiated nature of the process in which consciousness analyzes itself. This process is related to the nature of symbol in Jungian psychology and in the modeling field theory. I identify an essential connection among creativity, consciousness, the unconscious, and fuzziness, and attempt to delineate the boundaries of what is accessible today to the scientific method. Is it possible that mysteries of consciousness that are beyond rational understanding today are related to new physical phenomena, whose discovery will resolve the mysteries of matter related to the yet unexplained nature of quantum measurement and quantum gravity?

12.1 Phenomenology of Consciousness

12.1.1 Popular Conceptions and Misconceptions about Consciousness

Consciousness is an awareness or perception of inward psychological facts, a subjective experience of sensing, feelings or thoughts. This definition is taken from *Webster's Dictionary*. But a more detailed, rational analysis of consciousness has proven to be difficult. For a long time it seemed obvious that consciousness completely pervades our entire mental life, or at least its main aspects. Now we know that this idea is wrong, and the main reason for this misconception has been analyzed and understood: we are conscious only about what we are conscious of, and it is extremely difficult to notice anything else. In the beginning of his book on the historical evidence for the emergence of consciousness, Jaynes summarized eight popular concepts about consciousness, which he thought inadequate and useless. It is worth repeating them here briefly, because their popularity still has not vanished. Here they are. Consciousness is nothing but (1) a property of matter, (2) a property of living things, (3) a property of neural systems. These three "explanations" attempt to dismiss consciousness as an epiphenomenon, an unimportant quality of something else. They are useless because the problem is in explaining the relationships of consciousness to matter, to life, and to neural systems. Consciousness is not a simple correlate of any of these other "things," but has complicated relationships with them. These dismissals of consciousness are not very different from the behavioristic postulate that (4) there is no consciousness. But, of course, this conscious statement of behaviorists refutes itself. (5) A dualistic position is that consciousness belongs to the world of ideas and has nothing to do with the world of matter. But the scientific problem *is* in explaining the consciousness as a natural-science phenomenon, that is to relate consciousness and the material world. (6) Consciousness is learning. This was a position of associationism, popular among some scientists in the 1950s and 1960s. Associationism bounded consciousness together with experience and learning, but we know today that these are all different properties of mind. (7) Consciousness emerges in evolution. This is acceptable as a doctrine, but a lot of work is needed to find out what exactly emerged, when, and how it did. (8) Consciousness is a neural mechanism. Again, as a starting point it might be acceptable. But what is the mechanism, and the mechanism of what? Let us say we obtained the entire wiring diagram of all neurons in the brain, plus chemical properties of neurotransmitters, etc.; this still would not explain consciousness. Together with Jaynes we conclude that we need first to examine what consciousness is and what it is not. And I will add that a physical theory of consciousness would tie together the most important facts, with intuition and with mathematics.

Searle, in his book on mind (1982), begins the chapter on the structure of consciousness by mentioning subjects that he believes are crucial to consciousness, but are the least understood. The first subject is the relationship between consciousness and time: consciousness is not experienced as spatially extended, but is extended in time. The systematic differences between the phenomenological (perceived) time and real time are unexplained. The second subject is the relationships between the social and individual elements of consciousness, including the role of "other people" in the structure of our consciousness.

Then Searle lists 12 important properties of consciousness requiring scientific explanation: (1) finite modalities: consciousness is manifested in a strictly limited number of

modalities; these include the five senses, bodily sensations, thinking, and feelings (emotions); (2) unity: conscious states are parts of a unified sequence; Searle differentiates horizontal unity of a temporary sequence of events and vertical unity of simultaneous events; (3) intentionality: most conscious states are directed at something; we are conscious of something, and this "of" points to its intentionality; (4) subjective feelings: "what it feels like" more than anything else is responsible for the philosophical puzzlement of consciousness; (5) the connection between consciousness and intentionality: only conscious beings could be intentional; (6) structuredness of conscious experience: we are conscious of specific objects, events, concepts, and not of undifferentiated shapes etc.; (7) familiarity: we are conscious of objects (events, etc.) *as* specific types of familiar concepts; consciousness of something is consciousness of it as something; (8) overflow: conscious states refer beyond their immediate content; (9) center and periphery: consciousness is closely related to attention, a lot of things are at the periphery of consciousness and relatively few are at the center; (10) boundary conditions: the periphery of consciousness is located within boundaries of the situation; the boundaries (such as date, year, your name, country, when and what you will have for dinner) are important for the overall situatedness of consciousness, even though we are not necessarily conscious at every moment; (11) moods: are conscious, also not necessarily intentional; moods are pervasive "tones" or "colorings" of consciousness; (12) pleasure/unpleasure: is always a part of conscious states.

For one of these properties, (7) familiarity, Searle offers an explanation: "the categories have to exist prior to the experience, because they are the conditions of possibilities of having just these experiences." This becomes mathematically "obvious" when considering the mind as based on internal models: all our perceptions are possible only because of a priori model-categories. So we can neither perceive nor conceive, or be conscious of anything that is totally unfamiliar. In the following we will analyze the structure of consciousness and explain its many properties due to the internal model.

12.1.2 What Is Consciousness?

Our knowledge of consciousness is primarily of introspective origin. Understanding of consciousness requires differentiating conscious and unconscious psychic processes, so we need to understand what is psychic, what is unconscious, and what is consciousness. Our experiences can be divided into somatic and psychic. A will modifying instinctual reflexes indicates the presence of psyche, but not necessarily consciousness. Often we associate consciousness with a subjective perception of free will. Consciousness about somatic experiences comes against limits of unknown in the outer world; similarly, limits of consciousness about psychic experiences come against limits of unknown in the psyche, or unconscious. Roughly speaking, there are three conscious/unconscious levels of psychic contents: (1) contents that can be recalled and made conscious voluntarily (memories); (2) contents that are not under voluntary control, we know about them because they spontaneously irrupt into consciousness; and (3) contents inaccessible to consciousness. We know about the latter through scientific deductions.

Consciousness is not a simple phenomenon, but a complicated differentiated process. Jung differentiated four types of consciousness related to experiences of sensations, feelings, thoughts, and intuitions. In addition to these four psychic functions, consciousness is characterized by an attitude: introverted, concentrated mainly on the inner experience, or

extroverted, concentrated mainly on the outer experience. The interplay of various conscious and unconscious levels of psychic functions and attitudes results in a number of types of consciousness; interactions of these types with individual memories and experiences make consciousness dependent on the entire individual experience that makes an individual consciousness. An individual consciousness has a high degree of continuity and identity.

Consciousness is about something. In our theory of mind based on internal models, consciousness is about the internal model (of the environment, self, past, present, future plans and alternatives). Over the internal model, consciousness can direct attention at will. This conscious control of will is called free will. A subjective feeling of free will is a most cherished property of our psyche. Most of us feel that this is what makes us different from inanimate objects and simple forms of life. And this property is a most difficult one to explain rationally or to describe mathematically.[1] But let us see how far can we go toward understanding this phenomenon. We know that often raw percepts are not conscious. For example, in the visual system, we are conscious about the final processing stage, the integrated model, and unconscious about intermediate processing. We are unconscious about eye receptive fields; about details of visual perception of motion and color as far as it takes place in our brain separately from the main visual cortex, etc. These unconscious perceptions are illustrated in blindsight: a visual perception occurs, but a person is not conscious about it. In most cases, we are conscious only about the integrated scene, objects, etc.

It follows that internal models have conscious and unconscious parts that are accessible and inaccessible to consciousness. But what is responsible for the continuity and identity of consciousness? I propose a hypothesis that there is an internal model of the conscious parts of all other internal models. This a priori adaptive model is responsible for consciousness and the phenomenological properties of consciousness are due to properties of this model. Let us discuss what is known about the phenomenology of this internal model. Since Freud, a complex of psychological functions associated with this internal model is called Ego. Jung considered Ego to be based on a more general model or archetype of Self. Jungian archetypes are psychic structures (models) of a primordial origin, which are not accessible to consciousness in their entirety, but determine the structure of our psyche. In this way, archetypes are similar to other models, e.g., receptive fields of the retina are not consciously perceived, but determine the structure of visual perception. The Self archetype determines our phenomenological subjective perception of ourselves, and, in addition, structures our psyche in many different ways, which are yet far from completely understood. One of the most important phenomenological properties of Self is the perception of uniqueness and indivisibility.

Consciousness, to a significant extent, coincides with the conscious part of the archetype (internal model) of Self. A conscious part of Self belongs to Ego. Individuality as a total character distinguishing an individual from others is a main characteristics of Ego. Not all aspects of individuality are conscious, so, the relationships among the discussed models can be summarized to some extent, as

$$\text{Consciousness} \in \text{Individuality} \in \text{Ego} \in \text{Self} \in \text{Psyche}$$

Ego is a subject of free will. It possesses a free will inside consciousness. Free will is limited by laws of nature in the outer world and in the inner world by the unconscious aspects of

Self. Free will and intellectual aspects of Reason belong to consciousness, but not to the conscious and unconscious totality of the psyche.

Contemporary philosophers consider subjective nature of consciousness to be an impenetrable barrier to scientific investigation. Chalmers differentiated hard and easy questions about consciousness (Chalmers, 1994) as follows. Easy questions, which will be answered better and better, are concerned with brain mechanisms: which brain structures are responsible for consciousness? Hard questions, about which no progress can be expected, are concerned with the subjective nature of consciousness and qualia, subjective feelings associated with every conscious perception. Nagel described it dramatically with a question: "What it is like to be a bat?" (Nagel, 1974).[2] In the modeling field theory, the subjective nature of consciousness is not a mystery. It is explained due to the subjective nature of the internal model of which we are conscious. The subjectivity is the result of the MFT combining apriority and adaptivity, the unique genetic a priori structures of our psyche with our unique individual experiences. I consider the only hard questions about consciousness to be *free will and the nature of creativity*.

12.1.3 Consciousness of Bodhisattvas

The notion of emptiness takes a central and fundamental role in Buddhism. The emptiness of an object, in the terminology of MFT, means that its value for satisfaction of lower, bodily instincts is much smaller than its value for satisfaction of the higher instinct for learning. The consciousness of bodhisattva, writes Dalai Lama (1993), wonders at the emptiness of any object. This means that for bodhisattva any object is first of all a phenomenon available for our comprehension, its correspondence to our internal models excites the feeling of harmony, and its deviation from the models, its "disharmony," stimulates the process of adaptation, improving the internal models, and satisfying the learning instinct. In every moment of perception, bodhisattva consciousness is governed first of all by the learning instinct. The nature of emptiness in Buddhism is that the highest intention of every phenomenon is in its concept.

Reconciliation between the existential feeling of mystery and eternity and a practical need to live in the world of matter as a finite material being is a major purpose of all religions. In Buddhism, this reconciliation is achieved through the concept of emptiness. Boddhisattva consciousness that combines the concept and the will for improvement impressed Schopenhauer (1819) and stimulated his philosophical unification of the internal representation and will.

The affirmation that "Buddhas are always concentrated on emptiness" should not be confused with the emotional emptiness. The emotions that subsist in the flow of consciousness of bodhisattvas are related to the "higher" instinct for learning, and are not the "lower" emotions satisfying bodily instincts. Among lower "afflicting" emotions, according to Buddhists teachings, are attachment, aversion, and ignorance; these emotions are poisons of the mind causing mental and sometimes bodily illness. Contrary to the negative role of lower emotions, higher emotions are the neuronal mechanisms of will to perfect self. MFT offers a scientific way of understanding the Buddhists belief that when Buddhas meditate on the direct comprehension of emptiness, the higher emotions emanate from their minds: the consciousness of concepts, as the primary content of phenomena, leads to emotions satisfying the higher instinct.

12.1.4 Consciousness versus Unconscious

We know about psychic contents inaccessible to consciousness through scientific deductions. These deductions are of several origins: psychic, evolutionary, and neural. Deductive scientific analysis of the unconscious was initiated by Freud and Jung, who demonstrated the existence of the unconscious, examined its influence on mind, behavior, and consciousness, and investigated the nature of unconscious psychic processes. According to evolutionary arguments, our psyche has evolved from simpler organisms without consciousness and, possibly, from inanimate matter. Therefore we deduce that there have to be unconscious psychic structures, from which consciousness has evolved, and there have to be material structures in the brain supporting the process of the evolution of consciousness. Neural sciences discovered a number of neural structures supporting the consciousness, which are not accessible to consciousness directly. Nor are all the psychic processes supported by these structures conscious; during the past decades we have learned a great deal about relationships among neural structures and the conscious and unconscious psychic processes associated with them.

Unconscious contents can be classified into two general groups: personal and impersonal. Personal unconscious comprises life experiences that were forgotten or subliminally perceived, thought, or felt. Impersonal unconscious contents originate in the inherited possibilities of the psychic functioning in general. A significant part of these contents is common to all of humankind; Jung called these the *collective* unconscious. Deducing specific contents of and identifying archetypes or models in the collective unconscious is a difficult task. It is approached through analyses of dreams, myths, fairytales, and behavior of people. These analyses show that unconscious contents are not differentiated: archetypes of collective unconscious are models of psychic situations as a whole; they manifest as "fantasy-images" that are not related to visual perceptions and they are not differentiated into thoughts, sensings, or feelings. As we better understand the nature of archetypes in the human psyche, we should be able to develop similar type mathematical models for MFT-type systems.

Evolution of consciousness proceeds through differentiation of psychic functions and their contents. It seems that originally, at the dawn of consciousness, thoughts were no different than internal sensings or feelings. The original archaic state of consciousness is an undifferentiated identity. And only gradually, psychic functions differentiate and the faculty of Understanding acquires a large number of highly differentiated internal models-concepts, which make possible differentiated thinking. As far as categories of Understanding are concerned, we have made a good first step toward a mathematical description of the process of differentiation. MFT techniques for adapting a priori models to new objects and structures in the world and for estimating and increasing, as needed, the numbers of models describe the mathematics of this process.

In the process of differentiation, unconscious (or less-conscious) fuzzy concepts are adapted to the new data and become conscious low-fuzzy or crisp concepts. This process occurs every day, and may only take a moment of time when we perceive new objects and recognize new situations that are similar to those seen and recognized in the past. And this process may take days, years, or even many generations, when new concepts emerge, which make a profound impact on our consciousness. In the processes of perception, fuzzy a priori concept-models are usually unconscious, and only crisp concept-objects reach consciousness. In the processes of everyday conscious cognition, usually the concept-models

that are in our consciousness from past experience are modified to suit new situations. And when profoundly new concepts emerge, it might involve highly fuzzy primordial models-archetypes, that are made less fuzzy as they are brought into the consciousness in the process of adaptation.

Conscious models can be more easily communicated to other people than unconscious ones. And they can be adapted more quickly to the new circumstances as long as they retain a degree of fuzziness. These are the advantages of conscious models vs. unconscious ones, which determine their importance for the advancement of culture. The drawback of conscious models is that they gradually become crisp, and nonadaptive, and may lose their touch with the unconscious contents of the psyche. They might become shallow label-signs, a matter of seemingly arbitrary convention, designating nothing of importance to our lives, and they should be replaced by new, more meaningful concepts. This is the process of the cultural and conscious evolution.

12.1.5 Consciousness versus Emotions

Undifferentiated thinking is incapable of thinking apart from sensations, feelings, and intuitions. Similarly, undifferentiated feelings are mixed up with thoughts, sensations, and fantasies. For example, in neurosis, according to Freud, thinking and feelings are mixed up with sexual desires. Undifferentiated processes may "grab our guts," so that we are very well aware of them. Yet we have very little consciousness of what exactly goes on; our conscious self is not in control. On the contrary, in the case of differentiated thinking, we are conscious of many aspects of the situation, we are conscious of potential conflicts, and we can better adapt to the situation. Not only thoughts, but also other psychic functions acquire differentiated status. Jung considers four main psychic functions: thinking, feeling (emotion), intuition, and sensing. Thinking and feeling are rational and more easily accessible to consciousness and more easily attain highly differentiated status. Intuition and sensing are irrational and less accessible to consciousness and differentiation.

Emotions originate from evolutionary-old brain systems that control behavior essential for survival and reproduction. Therefore, emotional control of consciousness and undifferentiated, unconscious emotional influences on behavior are well known and documented. The opposite, conscious control of emotions, seems to be of relatively recent origins and is less understood. Have you ever fed seagulls on the beach? When a seagull sees a piece of bread thrown in his direction, he "cries wolf" in the seagull language. He fakes a danger-cry that was originally intended to warn a flock from danger, and he uses it intentionally and rationally (consciously?) to scare other birds, so that they do not compete for the piece of bread. Is this a rational fake of emotions? Is it a conscious one? If we do not acknowledge consciousness in a seagull, then we have to acknowledge that faking emotions is even older than consciousness.

Jung described an emotional type of psyche, in which highly rational intelligence is based on conscious differentiated emotions. The mathematical nature of the differentiation of emotions, however, is less understood than that of the differentiation of concepts. Let us remember the discussions in Chapters 2 and 10. Attempts to model emotions similarly to modeling concepts lead to a combinatorial explosion. This difficulty is avoided by using the mathematical desciption of emotions and concepts that follows the mechanisms elucidated in neural, psychological, and philosophical analyses; concepts are described by models,

and emotions are signals affecting the similarity measures (between the concept-models and structures in the input data). A mathematical description of the differentiated emotions is based on complicated structural similarity measures. Its highly differentiated development is based on the "emotional concepts." These high-value internal model-concepts are similar to instincts in that they affect similarity measures (between many other concept-models and input data). The differentiated similarity measures are affected not only by "lower" instinctual emotions, but also by "higher" intellectual emotions related to high-value "emotional" concepts. The high-value "emotional" concepts generate differentiated, conscious emotional signals and become as if differentiated instincts, subject to conscious control, learning, and will.

Unconscious undifferentiated emotions that are not under the control of free will are called affects. Affects can excersise powerful control over our psyche. But it would be wrong to assume that only undifferentiated unconscious emotions can be very strong. Very strong emotions can be highly differentiated, complicated, completely conscious, and under the control of free will. For example, Mazo (1994) describes lamenting, a custom observed in many areas of rural Russia, which consits of crying with words and melody on specific highly emotional occasions in personal life, such as marriage or a death in the family. It is not only a ritual, but also a conventional form of expressing certain affective, highly emotional states. Usually, people cannot lament at will so that a researcher can tape their laments. But there are professional lamenters, who are willing to lament on request; "they can even be interrupted in the middle of a lament and then carry on lamenting without discontinuity in the verbal content, mood, or emotional involvement." Of course, all of us are familiar with more refined versions of this type of behavior by actors in the movies or on stage, but the conscious, free will-controlled aspect is usually hidden.

Conscious parts of internal models are more differentiated, more adaptive, more accessible to will and Reason, and more amenable to future differentiation and adaptation. A predominant mode of consciousness determines the type of personality. Individuals of the thinking type (most of scientists) are most conscious of their faculty of Understanding, which operates with a large number of highly differentiated concepts. People of the thinking type could be relatively unconscious about their other psychic functions, and may not clearly differentiate among sensing, feeling, and intuition (say, internal feelings, as different from internal sensings and internal intuitions). Correspondingly, these less differentiated functions will operate with fewer differentiated models. The psychic significance of undifferentiated unconscious functions could be very high. Because they are fused with ancient affective mechanisms, they might exercise a strong control over the psyche, overpowering differentiated functions of recent origin. Most of the readers of this book, as well as scientists in general, are of thinking psychological types, consciously differentiating among a tremendous number of concepts and the relationships among them. We are much less conscious about our emotions. This does not mean emotions have less influence over us, just the opposite: less differentiated functions could be perceived as more "deep" and more "genuine." Jung goes to great length to demonstrate that the least differentiated experiences, by being involved with ancient affective psychic structures, could have the greatest grip over a person. This may lead to underappreciation of the most differentiated function and to overappreciation of a more primitive one. This phenomenon complicates scientific studies of consciousness, and, possibly, the lesser development of the science of emotions relative to conceptual thinking might be related to scientists being more conscious about their thinking.

Psychic experience has an especially high value if its conscious and unconscious aspects are in agreement with each other. A mathematical description of this fact in MFT is given as follows. A psychic experience is an excitation of an MFT–Kantian vortex involving models-concepts of Understanding and emotional similarity-Judgment. Consider now an incompletely differentiated model-concept, which is strongly connected to unconscious affective systems, e.g., it models highly desirable objects (say, sex partners). Judgment is responsible for recognizing objects that correspond to the model. Judgment is not an "impartial" similarity measure, but is affected by the value of the object, especially for poorly differentiated, highly desirable objects. Possibly two parallel aspects of Judgment are involved: one differentiated and conscious, which measures the similarity between the model and the object, and the other, less differentiated and unconscious, which is affected mostly by the desirability of the object. When both aspects of Judgment are in concert, the experience has a high value. An alternative situation, in which the unconscious aspect overpowers the conscious one, is colloquially called "loss of judgment due to affect."

An ability to coordinate and harmonize the conscious and unconscious aspects of Judgment is the essence of the emotional intellect. What are its mechanisms? The mechanisms of the emotional intellect seem to be in differentiation and bringing closer to the consciousness the affective, instinctual aspects of Judgment. Both Judgment and Understanding then become more adaptive. On the one hand, the affective aspects of Judgment are modified and harmonized with conscious and cultural aspects of models (instinctual desires are made more conscious and more cultural). And on the other hand, conscious and cultural requirements incorporated in the model-concept of the desired object are modified toward better correspondence to the instinctive desires (cultural and conscious requirements are modified to better meet our instinctual needs). In this process both our feelings and our internal representations are adapted and become more adequate.

Compared to differentiated thinking (Understanding), we understand much less about the processes of differentiation of other modalities of consciousness, such as intuition. Hopefully, the mathematics of emotional intellect described in Chapter 10 will lead to progress in this direction. I see ways toward developing mathematical methods of modeling the higher emotions in Kantian philosophy, in particular, in the Kantian discovery of the indirect nature of the aesthetical perception. Before Kant, beauty had been considered to be directly given to our feelings through a special gift of aesthetic sensitivity. Kant discovered the complex nature of aesthetic sensitivity and related it to the perception of potential for increasing knowledge. Sometimes enjoyment of beauty requires time and preparatory intellectual effort. For example, beauty of a scientific theory requires knowledge of the specific area of science. Yet even colloquial beauty or aesthetic judgment, according to Kant, is intimately related to an ability for learning. A mathematical theory of higher emotions can be discovered only by analyzing their specific a priori nature within the overall structure of the intelligence. Following Kant, beautiful is *truth,* perceived in estimative categories of the emotional. Or, in other words, beauty is an axiological existence of the category "truth." Truth here refers to the adequacy of internal representations.

Perception of beauty is a mechanism of evolution, adaptation, and survival. As with any other survival mechanism, there are mechanisms of camouflage, counterfeiting, and countercounterfeiting. This is especially true in sexual relationships (which is reflected in proverbs like this one: "the light is on, but nobody is home").

12.1.6 Why Is Consciousness Needed?

Why is there consciousness? Why would a feature such as consciousness appear in the process of evolution? The answer to this question seems clear: consciousness directs the will and results in a better adaptation. In simple situations, when only minimal adaptation is required, instinct alone is sufficient, and unconscious processes can efficiently allocate resources and will. However, in complex situations, when adaptation is complicated, various instincts might come to contradictions. Undifferentiated functions result in ambivalence and ambitendency; every position entails its own negation, leading to an inhibition. This inhibition cannot be resolved by an unconscious that does not differentiate among alternatives. Direction is impossible without differentiation. Consciousness is needed to resolve an instinctual impasse by suppressing some processes and allocating power to others. By differentiating alternatives, consciousness can direct a psychological function to a goal.

Totality and undividedness of consciousness are its most important adaptive properties needed to concentrate power on the most important goal at every moment. This is illustrated by clinical cases of divided consciousness and multiple personalities, resulting in maladaptation up to a complete loss of functionality. Simple consciousness needs only to operate with relatively few concepts. More and more differentiation is needed to be able to select more and more specific goals (by selective inhibition and excitation). The scientific quest is to explain the emergence of consciousness from unconscious in the process of evolution. Consciousness has emerged, driven by unconscious urges for improved adaptation. And the goal of consciousness is to improve understanding of what is not conscious, inside and outside of the psyche. Thus, the cause and the end of consciousness are to be found in the unconscious; it contains the causal mechanisms and goals of consciousness.

The above analysis of causes and goals of consciousness leads to an opposing question: why is unconscious needed? In case of living beings, including humans, it seems, the answer is simple: the unconscious is a product of our development from lower organizational forms of life and inanimate matter. And our evolution and personal goals are to increase consciousness. But is there a value of the unconscious for artificial systems?

12.1.7 Collective and Individual Consciousness

Conscious aspects of internal models are more adaptive than unconscious ones and they are more affected by personal experience, upbringing, and culture. In particular, consciousness leads to the development of the individual personality, that is, to the development and adaptation of internal models that differentiate people from their environment in many different ways, while preserving the personal identity by a synthesis of differentiated models into a coherent conscious whole. Jung called this process the individuation and considered it to be the most important task of personal spiritual development. Most of our conscious models are collective in nature in that they are conditioned by culture and only their relatively minor aspects are adaptive objects of the free will. I have long been puzzled by Descartes' error, by his denial of the soul to animals. Possibly, he had in mind the individual consciousness. In this, he would seem to be correct: individual consciousness is a human achievement; it is a historically new and still rare phenomenon among humans.

All of us believe that we have individual consciousness. To what extent is this the truth and to what extent is this our fancy? Before answering these questions we should

evaluate the diverse nature of collective or cultural types of consciousness. In particular, the culture promotes both individualistic and communal aspects of consciousness. There are significant differences among national and ethnic cultures in this regard. For example, North American culture promotes individualistic values to a larger extent than many other cultures. Latino-American or Russian cultures emphasize communal values to a much greater extent. Differences among cultures lead to differences in the predominant types of *collective* models of consciousness. Still, the development of the individual consciousness is a personal task.

The individualistic type of the North American collective consciousness should not be mistaken for the individual consciousness. And, it is not obvious which type of collective consciousness better prepares individuals for the development of the individual consciousness. Many immigrants find that American individualism exerts a tremendously liberating effect on the development of individual consciousness. But many Americans find the very same individualism to be tremendously stifling and feel liberated when abroad in a country with a predominantly communal consciousness. When the culture "imposes" a demand to be "an individual" on a child or a young person who is not ready yet for the task or who does not yet have individual consciousness, the result is often psychological compensation. The psychic energy is withdrawn from the too difficult task of building the individual consciousness and instead is redirected at finding communal values within the individualistic collective consciousness. Contrary to this, communal collective consciousness and closely knit extended families provide a nurturing element much needed for the development of individual consciousness, even while the communal values strongly discourage its development. Therefore, the development of individual consciousness remains a daring task in any society, and could be performed only by an individual.

Jaynes analyzed historical evidence for the emergence of individual consciousness. By studying historical events, Vedaic, Greek, and Babylonian epics, and the Bible, he came to the conclusion that the individual consciousness is a recent phenomenon, the emergence of which can be dated approximately around the second millennium BC. Individual consciousness is a learning process, the emergence of which paralleled cataclysms and catastrophes that interrupted a "hallucinatory mentality," or what I call early forms of collective consciousness. Archeological evidence suggests that about 9000 BC, rather abruptly, agriculture appeared in several places throughout the Near East, and during the next several thousand years, agriculture spread throughout the Near East, Anatolia, and Egypt. The great kingdoms of Ur and Egypt and their complex agricultural civilizations were based, according to Jaynes, on the "hallucinatory mentality" of the "bicameral mind," which perceived a priori collective concepts as direct communications from gods (Jaynes hypothesis of the "bicameral mind" refers to the concepts and thoughts generated in one hemisphere being perceived as god-given imperatives by the other). The stability of great kingdoms was supported by and was adequate for this type of consciousness. It is exemplified in the carving on the stele dated about 1750 BC that shows Hammurabi, a great king of Babylon, staying next to his god Marduk and attentively listening. The stele contains 282 god pronouncements or laws.

A stone altar of Tukulti-Ninurta I, king of Assyria, built just a few centuries later, about 1230 BC, shows a dramatically different picture. Tukulti is shown twice, first as he approaches the throne of his god, and second, as he kneels before the god's throne. No king before in history was ever shown kneeling. But what is even more remarkable, the

throne of god is empty. Jaynes' interpretation is that although the collective consciousness of Hammurabi reliably took the images and thoughts in its mind as communications from his god, the emerging subjectivity of Tukulti's consciousness destroyed this absolute belief in god, and "emptied" the god's throne. I would add that although Hammurabi's achievement was to formulate clearly and consciously what was previously unconscious and fuzzy in the collective consciousness, the more differentiated consciousness of Tukulti offers him more choices, so that he is uncertain, and he has to choose among his options.

Apparently, what happened in the mind of Tukulti became a collective condition: individuals in the society became conscious about multiple choices in their consciousness, and no longer obeyed the ancient traditions automatically. The law and order dwindled, and old methods of governing people became inadequate. Tukulti is the first of the cruel Assyrian tyrants, who conquered and ruled with unprecedented terror. His soldiers also had no taboos in their consciousness against massive cruelty. About the same time, between 1470 and 1230 BC, a volcanic eruption or a series of eruptions destroyed a large part of Aegean culture, possibly affecting the whole of the Mediterranean, Cyprus, the Nile delta, and the coast of Israel. This set off mass migrations and invasions destroying Hittite and Mycenean empires, and sending secondary waves of refugees and invaders throughout the Near East. The cause and effect relationship between geological cataclysms and emergence of consciousness, according to Jaynes, is not simple: some global cataclysms may speed up a need for the adaptation of consciousness; at the same time, changes in the paradigm of consciousness led to catastrophes destroying cultures and causing dislocations of peoples. Similar effects are known to have occured in America. Mayan civilizations periodically broke down, when quite abruptly the population left cities and went back into tribal living in jungles, sometimes with no trace of destruction or wars. In this century, an abrupt change of the consciousness paradigm occurred in Germany between 1933 and 1945 and, possibly, a change of consciousness paradigm is now occurring in Russia.

Maimonides, in his *Guide for the Perplexed*, analyzes the relationship between collective and individual consciousness. He was asked by his student, why did God, on the one hand, give Adam mind and free will, and on the other hand, forbid him to eat of the tree of knowledge? Did not God want Adam to use his mind? Before answering this question, Maimonides goes through three pages, denouncing his student for asking the most complex question of all, the question over which all the best minds of the entire humankind were raking their brains from time immemorial. And you dare to ask this question, between sex and morning coffee, without being prepared or being ready to comprehend the answer, continues the great theologist, to dissuade an unprepared one from reading the answer. And then he gives the answer for the one who is ready for it. God, according to Maimonides, gave Adam mind and free will to determine for himself what is good and what is bad. Instead, Adam took a "shortcut," he ate from the tree of knowledge of good and evil, and got ready-made rules of morals; he acquired collective consciousness. In conclusion, Maimonides explains that Adam's story describes our predicament. Although God expects us to think for ourselves, we cannot fulfill this expectation just by throwing away the rules of morals. Being the descendants of Adam, we acquired the collective consciousness, and our ability to "think for ourselves" rather than following moral rules is limited; developing individual consciousness is very difficult. The relationship between individual and collective consciousness as explained by Maimonides is essentially the same as the one given by Nietzsche.

What is the relationship between free will and the individual consciousness? Even if Ego is a product of culture and upbringing to a larger extent than of a personal free choice, it seems that free will exists within the conscious part of Ego. The individual consciousness is an ability to modify the structure of Ego by expanding the conscious parts of the internal models. It is "the icing on the cake," the highest achievement of human creativity.

12.1.8 Consciousness, Time, and Space

Consciousness if often described as inseparable from the perception of time. James (1890) described consciousness as "a stream flowing smoothly along, now eddying around, then flowing smoothly again." But this smoothness of flow might not always have been there. Contemporary perception of time as ordered and a smooth flow is an evolved feature of our psyche and, more specifically, consciousness. Various evidence points toward contemporaneous evolution of consciousness and time perception.

According to Jungian analysis, at the depth of unconscious are timeless archetypes.[3] Although figuratively calling them primordial "images," he emphasizes that they have no spatial structure either. Ancient archetype-models are atemporaneous and nonspatial; they possess no temporal or spatial characteristics, which possibly indicates that they have appeared before the perception of time and space. This corresponds to the Kantian conclusion that the perception of time and space cannot be a priori still a condition of any experience. Although at first the Kantian statement seems contradictory, the contradiction is resolved by considering the emergence of consciousness and perception of time as historically contemporaneous. The emergence of ordered time–space perception might possibly be described mathematically by the emergent locally linear behavior in globally nonlinear neural systems with competitive organization found by Grossberg. The emergence of conceptions of time and space was studied by psychologists (Lèvy-Bruhl, 1910) and linguists (Whorf, 1936), who indicated that some primitive tribes perceive time and space as locally ordered and unordered globally. In the region of space and time where the tribe lives today, events are ordered in space and time; the notions of *today* and *tomorrow*, *up to the* river and *beyond* the river, are clearly perceived and are not mixed up; however events separated by several generations or occurring beyond far mountains are not related by a continuous chain of causes and consequences, and their interrelations in time and space are broken and unordered. The perception of space and time characteristic of the primordial psyche, according to this interpretation, emerged from the continuous flow of stimuli, first as locally ordered and locally linear, and later as developed into globally ordered conceptions of time and space characteristic of the historic consciousness. Such an interpretation assumes a relatively modern origin of our contemporary a priori intuition of the globally ordered space and time, corresponding to a gradual emergence of neural structures. This is consistent with Jaynes' hypothesis of the recent origin of consciousness.

A relationship between consciousness and space–time perception is also indicated by an analysis of psychic processes during sleep. During rapid eye movement (REM) sleep eyes move, indicating spatial perceptions, and dreams during REM sleep contain significant conscious elements. During "deeper" non-REM sleep there are no eye movements (no spatial perception) and much less conscious content is remembered when awakened during non-REM sleep.

Internal subjective time could be very different from external "real-time." Especially interesting is that when solving scientific problems the mind is conscious, but not necessarily in space–time. Abstract scientific concepts are timeless and spaceless in their content (even if about space and time). This points toward creativity being related to an interaction between consciousness and the unconscious.

12.1.9 MFT and Searle Revisited

Let us summarize our discussion of the phenomenology of consciousness. Consciousness is an awareness or perception of inward psychological facts, a subjective experience of sensing, feelings, or thoughts. Consciousness directs the will and results in a better adaptation. In complicated situations, various instincts might encounter contradictions. Consciousness can resolve an instinctual impasse by suppressing some processes and allocating power to others. By differentiating alternatives, consciousness can direct a psychological function to a goal. Consciousness is a complicated differentiated phenomenon that is characterized (very roughly) by three levels: undifferentiated awareness, collective consciousness, and individual consciousness. Both collective and individual consciousness are characterized by multiple modalities or types of consciousness, that at the top level includes sensing, feeling, thinking, and intuition. Conscious parts of internal models are more differentiated, more adaptive, more accessible to will and reason, and more amenable to future differentiation and adaptation.

The following discussion is structured around questions and issues, which are emphasized by Searle in his books on mind (1980, 1992, 1997), as properties of consciousness requiring scientific explanation. According to Searle, the crucial but least understood properties of consciousness include the relationships between consciousness and time and between the social and individual elements of consciousness. According to our theory, consciousness is due to a specific internal model called Ego, and properties of consciousness are to be explained mathematically due to the properties of this model. Historical, linearly ordered conscious time is a relatively recent property of the Ego-model. Older time-perception mechanisms are related to various biological clocks (related, in turn, to metabolic rates of various bodily systems) that influence time perception on a short scale of minutes, hours, and days. On longer scales, especially on scales longer than an individual human life, the original time perception was not ordered.

The relationship between collective and individual consciousness, according to Searle, is the most neglected and puzzling topic. According to our analysis, the key toward understanding the roles of both types of consciousness is the differentiation between individual consciousness and individualistic values in collective consciousness. Even though most of us believe in possessing individual consciousness, the fact is that individual consciousness is a very recent achievement historically and still a very rare phenomenon. Most of our conscious models are collective in nature in that they are conditioned by culture and only their relatively minor aspects are adaptive subjects of the free will. The development of individual consciousness or individual personality is the result of adaptation of the conscious aspects of collective internal models. This process combines differentiation of a person from its environment in many different ways, while preserving the personal identity by a synthesis of differentiated models into a coherent conscious whole. Jung called this process individuation and considered it to be the most important task of personal spiritual

development. Individual consciousness is the ability to modify the structure of Ego by expanding the conscious parts of the internal models. It is related to free will and it is at the frontline of what can be rationally explained and mathematically modeled today.

Let us discuss other properties of consciousness, which according to Searle have not found scientific explanations in the past.

Unity. Conscious states are parts of a unified sequence and simultaneous events are perceived as unified into a coherent picture.

Finite Modalities. Consciousness is manifested in a strictly limited number of modalities; these include the five senses, bodily sensations, thinking, and feelings (emotions). These two properties are related and I discuss them together. Searle's unity is close to what Kant called "the transcendental unity of apperception." Again, this property is to be explained by the structural properties of the Ego-model and the process of its adaptation to the constantly changing world, in particular, the hierarchical organization and competition among submodels.

Let us begin the analysis of the relevant structures of the Ego-model from its preceding simpler forms. What is the initial state of consciousness: an undifferentiated unity (Jung) or a "booming, buzzing confusion" (James)? Or let us take a step back in evolutionary development and ask, what is the initial state of the preconscious psyche? Or let us move back even further toward the evolution of sensory systems and perception. When building a robot for a factory floor, why provide it with a sensor? Obviously, such an expensive thing as a sensor is needed to achieve specific goals: to sense the environment with the purpose of accomplishing specific tasks. Providing a robot with a sensor goes with an ability to utilize sensory data. (Why have sensors otherwise? I'll disregard a large number of expensive government programs building sensors without provisions for processing data. General Motors' factory floor robots are built more sensibly.) Similarly, in the process of evolution, sensory abilities emerge together with perception abilities. A natural evolution of sensory abilities cannot result in a "booming, buzzing confusion," but has to result in evolutionary advantageous abilities for avoiding danger, attaining food, etc. Initial perception abilities are limited to a few types of concept-objects (light–dark, warm–cold, edible–nonedible, dangerous–attractive, etc.) and are directly "wired" to proper actions. When perception functions evolve further, beyond immediate actions, it is through the development of complex internal models that unify individual object-models into a unified and flexible model of the world. Only at this point of possessing relatively complicated differentiated internal models composed of a large number of submodels, can an intelligent system experience a "booming, buzzing confusion," if it faces an entirely new type of environment. A primitive system is simply incapable of perceiving confusion: if its perceptions do not correspond to reality, it just does not survive without experiencing confusion. When a baby is born, it undergoes tremendous changes of environment, most likely without any conscious confusion. The original state of consciousness is undifferentiated unity. It possesses a single modality of primordial Self.

The initial unity of psyche limits the capabilities of an intelligent system, and further development proceeds through differentiation of psychic functions or modalities. This is true for robots as well as for biological systems, and this occurs in the process of evolution as well as in the process of individual growth. Not only do the internal models acquire more submodel-concepts, but the very nature of psychic functions differentiates. In particular,

differentiation between Understanding and Judgment or between thinking and feeling seems to be an important feature in the evolution of the animal kingdom and is crucial for the development of the human mind. With the development of consciousness, it also acquires this functional differentiation. About 2000 years ago humans became conscious of this differentiation. Bringing multiple psychic modalities into consciousness promises acceleration of the differentiation process and expansion of consciousness. Yet, 2000 years is but a moment in the genetic evolution, and we should not be surprised that we are still barely capable of consciously differentiating our own thoughts and feelings (if at all), and that we rarely use this differentiation for the betterment of our lives.

It is clear now that *finite modalities* of consciousness are determined by those psychic functions that reached consciousness. Consciousness is a recent phenomenon and many different psychical functions have not yet reached the level of consciousness. Jung differentiated four major modalities or functions of consciousness: thinking, feeling, intuition, and sensing. I challenge the readers to analyze their consciousness and identify these four modalities. It is a daring task! MFT provides the mathematical description of thinking and feeling (as it relates to consciousness and higher mental abilities). There are no conceptual difficulties in developing mathematical descriptions of "lower" feelings. But what is the mathematics of intuition?

Intentionality. Intentionality is a property of referring to something else. Most conscious states are directed at something; we are conscious of something, and this "of" points to the intentionality of consciousness. In everyday life, when we hear an opinion we do not just collate it in our memory and relate it to other opinions (like a pseudoscientist in a comedy); this would not lead very far. We wish to know what the aims and intentions associated with this opinion are. Often, we perceive the intent of what is said better then specific words, even if the words are chosen to disguise the intent behind causal reasoning. The desire to know and the ability to perceive the goal indicate that in psyche, the final standpoint is more important than the causal one. This intentionality of psyche was emphasized even by Aristotle in his discussions of the end cause of the Forms of mind. The consciousness is fundamentally intentional.

The intentional property of consciousness led many philosophers during the past decades to believe that intentionality is a unique and most important characteristic of consciousness: according to Searle (1992), only conscious beings could be intentional. These recent attempts to interpret intentionality as uniquely characteristic of consciousness seem misdirected. Intentionality is a fundamental property of our internal models: every one of our internal models has evolved with the intent or purpose of recognizing a particular type of signal (events, messages, concepts) and act accordingly (e.g., send a recognition message to other parts of brain and other models, including motor-control models). The first one to offer this explanation of the intentionality of mind was Aristotle, who argued that intentional states should be explained through the a priori contents of mind. [He called intentional states the "end causes of Forms" and he called a priori contents of mind the "Formal causes," that is the a priori properties of Forms (*Metaphysics*).]

Within a living system everything is intentional; intentionality is the property of life. Also, *every* concept or object that could be recognized by an artificial intelligence system is intentional: if it is recognized, it is always with the intent to accomplish something (otherwise, the very concept of recognition is meaningless[4]). General concepts belong to

the realm of pure spirit, a part of mind that is mathematically described in MFT by internal models. Specific individual objects belong not to the outer world of matter or manifold, but to the phenomenal world, the realm of interaction between the mind and matter. Thus, every object is intentional. It is important to differentiate this statement from a philosophical position of *pan-psychism,* which assumes that the matter itself, in all its forms, has a degree of psyche or spirit as its fundamental property. Pan-psychism does not really explain matter or psyche. This is why Descartes exorcised spirit from the world of matter. To a significant degree, pan-psychism is a result of failure to recognize the difference between the world of matter (manifold) and the world of phenomena.

This analysis of intentionality of the phenomenal world is applicable to cultural phenomena as well. Every cultural concept and every man-made object are intentional because they emerged, or were created (consciously or otherwise) with a specific intent (or purpose). The intentionality of objects has two aspects: their "high intellectual intention" is to correspond to the world (to be efficient for the analysis of the manifold and ultimately for survival) and their "low intellectual intention" is to be used for the intended utilitarian or instinctive purposes [e.g., a table in my kitchen is not just a thing in itself, but an intentional concept-object; its "high intellectual intention" is to recognize the table as a part of material world and use it for building a coherent picture of the world in my mind (my internal model); and its "low intellectual intention" is to use the table-matter appropriately for sitting and eating, etc.; in this regard the table is an external representation of the concept "table"].

Is there any specific relationship between consciousness and intentionality? If anything, it is just the opposite of Searle's hypothesis of one implying the other. Affective, subconscious, "low-intellectual-level" emotional responses are concerned with immediate survival and utilitarian goals, and therefore are intentional in a most straightforward way. A high-level consciousness is not concerned with immediate survival, but with the overall picture of the world, with knowledge and beauty; it can afford to be "impartial," abstract, and less intentional than the rest of the psyche. A mathematical description of this aspect of intentionality is given in MFT by the structure of purposes of its submodels, the intent of which is learning and adaptivity toward maximizing the total similarity between the world and its model. The highest creative aspect of individual consciousness and the ability to perceive the beautiful and sublime are intentional without any specific, lower level utilitarian goal; they are intentional toward self-realization, toward the future self beyond the current self.

Subjective Feelings. Subjective feelings, or what philosophers call qualia, "what it feels like," according to Searle are more than anything else responsible for the philosophical puzzlement of consciousness. Subjective feelings accompanying perception should be analyzed within a context of a two-way interaction between individual sensations and the internal models. The model, due to its adaptive property, includes inherited structures, individual development, and the remains of the entire accumulated experience up to the moment of a particular individual perception. The properties of qualia become even more complicated when one accounts for conscious as well as unconscious aspects of the internal model. The subjectivity of qualia thus does not seem to be a mystery beyond mathematical description, just the opposite; within the theory of consciousness based on adaptive internal model, subjectivity is a natural property within the mathematical theory.

Structuredness of Conscious Experience. We are conscious of specific objects, events, and concepts, and not of undifferentiated shapes etc. We are conscious of our internal model (or

more accurately of a portion of it). We are conscious of the manifold of the outer world only through the corresponding concept (model) in our mind. But we are not conscious of the outer world in any "direct" way. The subjects of our perceptions are phenomena, which are our submodels in their interaction with the world. And our consciousness is even removed from the "raw perceptions": we are conscious most of the time only of the "high-level" submodels, when they are combined in our mind into a unified model of the world after several layers of processing. This is best illustrated in the so-called "blind vision," when a person sees without being conscious of it. Complicated multilevel processing of visual perceptions before they reach consciousness is well documented (Zeki, 1996).

The structure of consciousness determined by the internal model explains several features discussed by Searle.

Familiarity. We are conscious of objects (events, etc.) as specific types of familiar categories. Consciousness of something is consciousness of it as something familiar: "the categories have to exist prior to the experience, because they are the conditions of possibilities of having just these experiences." Since we are conscious of the internal model, which is in our mind in some a priori form before perception, we cannot perceive or be conscious of "complete novelty."

Overflow. Conscious states refer beyond their immediate content. This is because the submodels used to perceive the immediate contents are interrelated with a large number of other submodels of the internal model. And the submodels reach consciousness only within the unified internal model, which always refers beyond the immediate content. This property is related to the functioning of attention, presence of the *center, periphery,* and *boundary conditions* of consciousness. A lot of things are at the periphery of consciousness and relatively few are at the center. The very periphery of consciousness is situated within the boundaries (such as date, year, your name, country, etc.), which are important and easily accessible to consciousness, even though they are not necessarily conscious at every moment.

We have discussed a number of properties of consciousness that were identified by Searle as requiring scientific explanation, and we discussed the explanation of these properties within the modeling field theory. This discussion is a step toward a future detailed mathematical description of consciousness, toward creating robots with the elements of consciousness. I would like to conclude this discussion by commenting on what Searle calls the *pleasure/unpleasure dimension* of conscious states. A mathematical explanation for the most of the above properties of consciousness is due to the properties of the internal models. The pleasure/unpleasure dimension (as far as it refers to something beyond the "lower emotions" and satisfaction of physiologically related instincts) is due to the faculty of judgment. Every minute conscious perception of the objects in the outer world as well as internal thoughts is due to the ability of judgment to recognize that a subset of the manifold corresponds to a particular concept-submodel. This correspondence causes the pleasure dimension of consciousness, and absence of the correspondence causes unpleasure or pain. In a hierarchically organized consciousness, thoughts and concepts of mind at every level are "perceived" by the next higher level in a similar way: through the judgment that a lower level concept is an instant of a higher level concept. To the extent that our consciousness is not hierarchical but also heterarchical, our discussion equally refers to parallel modules of the internal model as well as to the lower higher levels. This provides a mathematical foundation for describing pleasure/unpleasure as a ubiquitous part of conscious states.

12.1.10 Neural Structures of Consciousness

12.1.10.1 Higher Brain Functions

The anatomy of the brain reveals a mechanism of tremendous complexity. An architectural organization of the brain is not very ordered: it is a heterarchy of various modules, evolved through hundreds of millions of years from disparate sources and for disparate purposes. Within some specialized areas or modules of the brain there is a significant degree of order, e.g., the retina or primary visual cortex exhibits significant architectural coherency. Many individual brain modules exhibit ordered hierarchical organization. But even within such hierarchically organized modules, information processing is not strictly hierarchical; loops of feedback connections are often present and are important for information processing, as revealed by psychological experiments. Therefore, ordered perception is not entirely due to ordered architecture, but to internal models, organizing the information. Even more so, higher psychic functions of more recent origins, such as differentiated consciousness and conscious control of emotions, might not necessarily be due to a highly ordered architecture, but rather to the remarkable adaptation capabilities of the mind. The adaptation capabilities are due to specific properties of a priori models. The a priori models are encoded in the brain (some are hardwired, others are learned), but this encoding is not necessarily related to the global brain architecture in a straightforward way. So the order of our mind processes could have more of a psychological nature, related to fine details of adaptive neural connections, rather than to the global architecture of the brain.

The cerebral cortex is a most distinctively human part of the brain, in that the area and complexity of the cerebral cortex progress from lower to higher animals. The area of a human cortex is about 2500 cm^2; its volume is about 300 cm^3 and it contains about 10^9 neurons. Individual neurons may have up to hundreds of thousands synapses. The cerebral cortex of humans is exceeded by that of the whales,[5] but the human cortex is probably the most differentiated in terms of the number of distinct subareas. The most recent evolutionary part of the cortex is called the neocortex, and it is present in all mammals. In submammalian brains, some parts of the cortex are present, but neocortex is very small if present at all. The human neocortex is responsible for higher brain functions. This is known from large amounts of various types of data, such as electroencephalograms and magnetic resonance imaging data collected in conjunction with psychological experiments, studies of patients with damaged brains, etc. Specific functions have been identified with specific areas of the cortex, such as several functionally different areas of visual cortex, speech production and speech understanding areas, etc.

12.1.10.2 Seat of Consciousness

Where in the brain is the seat of consciousness? Is there a specific single module responsible for consciousness? A candidate area often considered in this regard is the nucleus reticularis thalami (NRT), a thin layer of cells lying around the thalamus along its surface. The NRT and thalamus are ancient parts of the brain, located next to the brainstem. The thalamus is interposed between the brainstem and forebrain; it is a "relay station" that relays sensory and motor information to the cortex. In the waking state, it transmits incoming messages with high fidelity. In deep sleep, the transmission is blocked. NRT neurons do not directly send their axons to the cortex, but virtually all connections between the thalamus and cortex pass through the NRT and have synaptic contacts with it. In this way the NRT controls

interactions between sensory systems and cerebral cortex. The NRT is the major site of attentional control over cortical activity. Damage to NRT leads to loss of consciousness. These as well as other more detailed considerations led many researchers to implicate NRT in consciousness.

Involvement of ancient parts of the brain in consciousness supports the hypothesis that consciousness is of an ancient origin. This corresponds to a layperson's perception that among mammals and possibly even lower animals we can easily observe the difference between conscious and unconscious states. In the past, our humanness was often identified with consciousness without differentiated analysis of what it is, and the issue of the recent or ancient origin of consciousness was a subject of many erroneous opinions and conclusions. In view of the preceding discussion in this chapter of the various levels of consciousness, it should be clear that human consciousness is very different from the simple awareness that we perceive in lower animals. According to the modeling field theory, consciousness is the consciousness about the internal model, so the difference in levels of consciousness is explained mathematically, in part, by differences in the complexity of internal models.

Taylor (1994a,b) explains consciousness as a winner-takes-all competition among cortical centers. The mechanism is the inhibitory connections in NRT. According to Taylor, consciousness resides in an NRT–thalamus–cortex system of neuronal connections. Episodic memories may affect consciousness through connections between hippocampus (containing memories) and thalamus. Taylor postulates a winner-takes-all mechanism in order to explain uniqueness, indivisibility, and concentration of consciousness. Although winner-takes-all explains conscious allocation of attention, it does not explain the maintenance of the coherent picture of the world and self characteristic of consciousness. A more complicated mechanism is needed, such as competition within a heterohierarchical model structure of MFT.

Other brain regions involved in consciousness include the hippocampus, which is responsible for memory formation and processes all inputs from sensory and association cortex for memory storage, and the amygdala, which is involved in emotions (feelings of rewards and punishments) and is responsible for affective evaluation of sensory stimuli. The frontal lobes coordinate selective attention activation of the hippocampal and amygdalar systems.

The amygdala has an important role in integration and control of emotional behavior. It receives multimodal sensory inputs from both external and internal environments and is involved with integrating them together with previous experience and proper emotional response. Through extensive axonal projections to many brain parts the amygdala controls emotional behavior. The amygdaloid complex is a specialized cortical region located in the temporal lobe. It is closely related to the olfactory complex and receives substantial inputs from all sensory system. It has extensive axonal projections to many brain parts, including the thalamus, brainstem, hippocampal formation, prefrontal cortex, and several other cortical areas. It is involved in cognitive functions, including memory processing.

The hippocampal formation plays an important role in formation and storage of memory, most likely together with several other brain regions. Together with the amygdala, it may be involved in storing memories with emotional content. It receives input from all major sensory areas of the neocortex through the entorhinal cortex, which can be thought of as a summary of the environmental information.

The amygdala and hippocampus are parts of the limbic system, which is involved in elaboration and expression of emotions. It is a phylogenetically ancient ring of medial cortex, including olfactory cortex, hippocampal formation, and several subcortical regions, which share direct cortical connections, including the amygdala and parts of the thalamus. The limbic system forms a highly interconnected structure, lying between the neocortical association region and hypothalamus (an ancient part of thalamus). It may serve as the gateway for neocortical cognitive influences on hypothalamic mechanisms of motivation and emotion.

Levine places the executive function associated with high-level consciousness in the prefrontal cortex. The executive function coordinates and integrates plans of actions, sensory signals from the environment, and motivational signals from the organism. It performs cognitive–emotional integration, links events through time, and joins working memory representations. It is not needed for routine sensorimotor tasks or ordinary memory (e.g., conscious perceptions take longer than automatic reaction time). With unambiguous signals to perform specific actions, the executive is inactive. When signals are weak or ambiguous, the executive coordinates lower level systems. Current internal states, mediated in part by the amygdala (that adds emotional coloring), are combined in the executive with the current environment, mediated in part by the hippocampus (memories), in order to guide actions. In patients with prefrontal damage, affect could be as strong as in a normal person, while not guiding actions. Also, normal restraints on emotional expression could be impaired. This indicates that the prefrontal cortex is involved in higher level, possibly conscious control over emotions. Damage to the prefrontal cortex impairs flexibility of adaptation, even though patients with such damage perform some specific learning tasks as well as normals.

12.1.10.3 Kant–MFT Neural Circuits

Let us consider candidate brain areas containing Kant–MFT-type neural circuits. A detailed investigation of this topic would be a subject of future research and a separate book. Here, I just briefly list candidate areas without going into detail. It is important to keep in mind during the following discussion that cortex, possibly, contains "copies" of subcortical regions, so that functions of ancient brain areas can be modified and influenced by higher brain functions located in the cortex. Internal models (Understanding) are first associated with memories. There are several types of memories: episodic, short-term memory, long-term memory, working memory, explicit declarative memory, and implicit procedural memory. The structures involved in memory formation are the cerebellum, hippocampus, amygdala, and cerebral cortex. The cerebellum is involved in learning and storing motor-control memories. These memories are considered to be implicit, procedural, and nonaccessible to consciousness directly. Also, skill memories are located in the frontal lobe/basal ganglia. Explicit, accessible to consciousness long-term memories are associated with the hippocampal complex. Language production and understanding models are located correspondingly in Broca's and Wernicke's areas of the cortex (both are in the left hemisphere in most people). It is interesting that these areas affect both spoken and written language in a similar way, so that the models are linguistic, rather than speech models.

Similarity measures (Judgment) seem first associated with the amygdalar and other limbic systems responsible for emotions, and possibly with other cortical regions. Kant–MFT adaptive symbol-formation processes related to visual recognition, learning, and eye movements (visually initiated motor sequences) are located in frontal eye fields and

supplementary eye fields that are located in proximity to the supplementary motor cortex. According to Levine (1996), in making choices, "both rational biases through the hippocampal-dorsolateral system (models, memories, Understanding) and emotional biases through the orbital-amygdalar system (similarity, Judgment) are more likely to operate simultaneously in real time, and both have conscious and unconscious components. Also, unconscious effects can influence the processing done by the conscious system, and vice versa." It is interesting to mention that there is a synchrony of neural firing associated with conscious perceptions; it might indicate that similarity measures of the conscious aspects of models, $\mathrm{ll}(n|k)$ and $f(k|n)$, are encoded in the brain by the degree of synchronicity of the n-signals (data, input) and k-signals (internal models).

12.1.10.4 Emotions and Consciousness

Conscious control over emotions is limited by the architecture of the brain structures implicated in emotions and consciousness. NRT is the major site of attentional control over cortical activity implicated in high-level consciousness. But NRT does not have a similar influence over the limbic system, which is implicated in emotions. The limbic structures (amygdala and Papez circuit) project their axons on the thalamus, which in turn projects on NRT, giving an "emotional color" to consciousness. But there is no direct return connections from the NRT to the limbic system. High-level conscious control over emotions may be effected in the prefrontal cortex, which integrates lower level systems and seems to contain "copies" of subcortical circuits. There are strong reciprocal connections between the orbital region of the prefrontal cortex and amygdala. Functions of the prefrontal circuits and connections are not sufficiently known and remain the subject of speculation. The fact that high-level conscious control over emotions is difficult for many people and that it could be learned supports the hypothesis that it is of a relatively recent origin and, thus, located in recent regions of the brain, that is in the cortex.

12.2 Physics of Spiritual Substance: Future Directions

12.2.1 Path to Understanding

The way toward understanding, in general, is through the cycles of differentiation and synthesis. Differentiation of our psychic processes into thinking and feeling, in particular, is a great discovery facilitating understanding of mind. Origins of this differentiation could be traced back more than 1800 years. One of the first descriptions of human psychological types was given by Galen (II AD). He described the famous classification of human psychological types in four temperaments: the sanguine, the phlegmatic, the choleric, and the melancholic. Using a more standard psychological terminology of today, temperament is a degree of affectivity or emotionality. This classification, which does not recognize thinking as different from emotions, has persisted for almost 2000 years, despite the fact that the differentiation between thinking and emotional types can be traced to Gnostic psychology contemporary to Galen. Gnostic psychology described three psychic types: pneumatikoi, psychikoi, and hylikoi, which could be related to thinking, feeling, and sensing. Affectivity is a psychological characteristics, which is most easily perceived from the outside. Other psychological functions are more difficult to perceive and for many, they are still barely

perceived subtleties. Notwithstanding, the affective side of the personality is considered today as secondary, as in: "I can not be held responsible for YY, because I did YY in a state of affect." Thus, even though not without ambivalence, we identify ourselves primarily with consciousness and free will, which are mostly associated with thinking.

On the one hand the process of differentiation between thinking and emotion is still in progress and not quite completed at the individual level. On the other hand, at the level of collective consciousness, in the western cultural tradition, this differentiation for a long time was perceived as an opposition between thinking and feeling. Often, we tend to forget that this opposition is a highly abstract, refined, top-level view of our psyche. Actual human processes of intellection are complicated interwoven interactions, vortices of multiple neural processes, involving various parts of the brain and various conscious, unconscious, and instinctual levels. Synthesis of differentiated psychic functions with the goal to understand our psyche as a whole is of a relatively recent origin.

The path of differentiation and synthesis is followed by philosophers and researchers of the mind. Aristotle, Kant, Jung, and other philosophers analyzed our intellectual and psychic processes by differentiating them into a number of differentiated concepts. And their attempts to understand the intellect in its wholeness is a process of synthesis following the differentiation. In particular, Aristotelian logic is one highly differentiated aspect of our mind. But a historic view of rationality as limited to Aristotelian–Gödelian logic is too narrow and is being rejected today. In the process of synthesis, a new understanding of rationality emerges as a hierarchical goal-directed functioning, which involves internal and external actions: actions within the mind of an intelligent system and into the outside world. And actions require an estimative functions, that is emotions. Thus, rationality has to include emotions. This understanding, initiated by Kant, reemerges today concurrently in multiple fields dealing with phenomena of intelligence: in philosophy, cognitive sciences, art and art criticism, education, mathematics, and engineering. Chapter 10 described the mathematical theory of MFT combining understanding with emotions.

Physical understanding is a specific intuition about the world. Possibly, this is what Newton meant when he said that he does not invent hypothesis about the world, but he knows how the world is made up. Contemporary physics is mostly concerned with the external world. But it was not always this way. Newton biographers indicate that he was disappointed that his physics was limited only to the external material world. During the nineteenth century, great physicists (including Humboldt and Maxwell) explored the physics of the inner world, however, the necessary mathematics of nonlinear, nonlocal field theory did not exist (Grossberg). Ours is a fascinating time of making of the physics of the inner world, and more generally, the physics of spiritual substance. A step toward physical understanding of the intellect including its conscious and emotional aspects is attempted in this book. Needless to say, only a preliminary sketch of some aspects of a future theory is possible at this time. Difficulties along this way are not less than those faced by Newton, who needed to develop both physical intuition and the mathematical apparatus to explain the data collected by Copernicus and other astronomers. The physical theory of spiritual substance, which encompasses consciousness and emotions, has to develop its intuition, mathematics, and the data. The additional difficulty is related to the nature of data: although everyone has an immediate intuition about consciousness and emotions, there is a widespread misunderstanding in scientific circles about the nature of these data.

A level of physical idealization should be appropriately chosen: if Newton started with quantum properties of space, we might still not have a theory of gravity. Were he not aware of the astronomers' observations, he would not have come up with the gravitation theory either. Today, the challenge is in selecting a proper level of physical idealization in the ocean of neurophysiological and psychological data, which is still grossly incomplete for creating a unified theory of mind and brain. Simplifications and idealizations are necessary for a preliminary exploration of such a wide subject. Study of mind proceeds today at many levels of physical idealization: for example, Grossberg emphasizes ties between physical intuition and mathematical modeling of neural structures. My analysis emphasizes ties between physical intuition and mathematical modeling of philosophical conceptions of mind.

12.2.2 Physical Nature of Symbol and the Emergence of Consciousness

Cosmogonic myths, according to Jung, are symbols of the coming of differentiated consciousness (*Aion*, p. 148). The consciousness was emerging in the process of ordering the percepts out of their original unordered chaos. This explains why original myths describe creation of ordered cosmos out of chaos. The ordering had to be achieved based only on what was "available" to the psyche, or by "analogies" with previously established more primitive a priori models, archetypes of Mother, Father, archetypal family relationships, etc. The origin of these archetypal models goes back to mechanisms of imprinting observed today in animal's newborns and disappears in the history of evolution. At some point in history, our ancestors started using these archetypal models for differentiation and organization of perceptions about the outer world on a more abstract level than that of animal psyche.

The vortexes of interacting archetype-models and percepts that appeared in the ancient psyche are ancient symbols, traces of which had been later recorded in ancient myths. The emergence of individual differentiated consciousness, according to Jaynes, occurred during the second millennia BC. Collective differentiated consciousness, according to Jaynes, was developed from about the ninth to third millennia BC. By this time, primordial archetypes existed for millions to hundreds of millions of years, and became "hardwired" genetically inherited a priori models that made psyche possible.

The process of emergence of differentiated consciousness has been traced in anthropological research comparing consciousness among peoples and cultures. Levy-Bruhl describes a psychic phenomena observed in primitive peoples that he called participation mystique. It is a psychological connection of a person with an object, in which subjects cannot clearly distinguish themself from the object. This partial identity with an object is of an a priori origin and is a property of undifferentiated consciousness. Participation mystique is a vestige of this primitive state of psyche. Although it was first observed among primitives, it occurs frequently among civilized peoples. Identification with an object impairs cognition and interferes with adaptation and survival. In this regard Jung quotes a case (observed in a primitive tribe) of a father killing his beloved young son when in a rage over a failed fishing trip. We do not have to live among primitive tribes to observe this type behavior. It is unfortunately still too common among civilized peoples.

Identification with an object in the realm of affect could be especially devastating, but even less complete identification at the level of concepts interferes with cognition. Adaptivity requires development of the differentiated consciousness and this is achieved

through symbol. In psychological terms, symbol "draws libido away from the object, devalues it, and bestows the surplus libido on the subject." In mathematical terms, symbol is a loop of interactions among percepts-cognitions, internal models, and feelings. Symbol results in adaptation of the internal model, which becomes more differentiated, and, in turn, leads to a more differentiated and beneficial direction of psychic energy.

This basic nature of symbol as a process of differentiation and adaptation is similar in everyday perception and in the process of evolution of consciousness. In everyday perception, the amount of adaptation is relatively small and a symbol-process quickly converges to a definite conscious concept-sign (leading to a behavioral or mental response in terms of previously learned concepts). In the historical evolution of consciousness leading to changes of cultural paradigms, symbol-processes may persist for thousands of years, which requires cultural, external representations of models. The symbol-process persists through cycles of internalization and projection by many individuals over millennia. We hypothesize that similar symbol processes sustained over much longer periods lead to changes of genetically inherited archetypes. The development of the physical theory of genetic evolution of consciousness would require verification of this hypothesis, clarification of the nature of interaction of MFT loops with genetic evolution, and further development of concepts of evolution.

12.2.3 Nature of Free Will and Creativity

12.2.3.1 Cat Named Schopenhauer

People who love cats are not surprised to hear that some cats are geniuses. I do not belong to this group of people by birth. I am convinced that people are endowed with more intellectual abilities than cats. But Schopenhauer surprised me and made me think again about the nature of the consciousness and creativity. He was a neighbor's cat, who at some point in his life decided to come live with us. It was a surprising decision, particularly because he was admired in his own home by his human masters as well as by his older dog-friend who brought him up and took good care of him. And because, as mentioned above, I am not a natural lover of cats, so he was not universally welcome at our home. But he made me like him. How did he do it?—I cannot explain. Why did he come to live with us? For a while, Schopenhauer was going back and forth between his old home and ours. His owner was asking my wife what kind of food we buy for him etc. and was trying to lure him back in every way. To this day I think that he came to us because he enjoyed the philosophical atmosphere at our home. He had a strong will. And my wife named him Schopenhauer, because he always got his way.

For a while Schopenhauer kept me in quandary about the nature of consciousness. Our social life is quite extensive and I have many friends. With friends you can enjoy playing tennis, going to movies, discussing books, and talking about philosophy. None of this is possible with a cat, so why did I feel that Schopenhauer is a genius, while not many people fit into this category? What does it tell me about the nature of genius? One day, unexpectedly, the answer to this question was revealed to me. My wife and I were walking down the street by Schopenhauer's old home. At the time Schopenhauer was still going back and forth between the two homes, and he happened to be in the frontyard of his old home. We called him. He came up, but not very close. He recognized us as distinct from other passersby, but he definitely did not quite recognize us as those people that he knew in our home. He

could not perceive us separately from surroundings as clearly as any human recognizes his friends. His consciousness was not sufficiently differentiated for this. It became clear to me that there are relatively few differentiated concepts-submodels in his internal model, just those that are vitally important for his survival. Probably there are models for mice, birds, and dogs, a number of models for odors and smells, a few models for people, and a few models for environments, still their degree of differentiation is much less than that of our human models.

But within this relatively narrow a priori internal model, Schopenhauer was free. He could even modify his a priori model, move its boundaries a little to suit his needs, such as changing his home and making others like him, even those who are not inclined to, such as myself, or an unfriendly dog. A few people can do this. Genetically and culturally people inherit a tremendously wide, differentiated internal model composed of hundreds of thousands or millions of submodels. But most could barely find their way around just a few familiar corners. Not many can fully use their inheritance: people "bury their talents." And very few can modify their a priori models according to their needs. Those who can make another human being like them, even against her or his inclinations, become great leaders. They lead to changes of our internal models through individual impressions or by cultural means.

One day Schopenhauer disappeared. We mourned him for a long time. During this time, my wife told me that she thought Schopenhauer was a female. I doubted it, but there was no way to verify this.

12.2.3.2 Creativity and Differentiation

An ultimate realization of free will is creativity. Modeling field theory of mind leads to a conclusion that creativity consists in the development and expansion of the conscious parts of internal models. It includes differentiation and synthesis. Differentiation is first the differentiation of subject from object, and next the differentiation of psychic functions, such as concepts, from feelings. Differentiation also takes place within each psychic function: differentiation of thinking is achieved by developing differentiated internal models. MFT contains mathematical description of some aspects of this process: generation of new models and reducing fuzziness of models in the process of learning; still we should better understand the role of evolutionary computations and genetic algorithms in the process of differentiation. Differentiation of feelings and other psychic functions is less understood. I think differentiation of feelings proceeds through the development of differentiated models of emotional concepts, especially those involving relationships.

The nature of creativity as well as the nature of beauty gradually change in the course of evolution. At the dawn of consciousness, subject–object differentiation required unusual creative powers. We see traces of this process in ancient myths. Culturally, as a part of the collective consciousness, this process is long accomplished, even though it is not necessarily complete in every individual psyche. Every individual has to perform the creative acts of differentiating himself or herself from the world, from his or her mother, from the object of sexual desire, etc. Most often, this task is no longer concentrated on the subject–object differentiation, but it has evolved toward a more fine differentiation of psychic functions. During most of the past 2000 years, the creative power concentrated on differentiation of thinking and feelings. Philosophers and scientists addressed concepts of understanding, while poets and artists developed differentiated concepts of feelings. Plato

and Aristotle developed concepts of thoughts abstracted from feelings. The philosophy of ancient Greece gave rise to a popular notion of a philosopher as aloof and removed from real life. Poets and artists aimed at impressing viewers and listeners by appealing to affective sides of psyche. During the past several hundred years this process is changing toward synthesis.[6] According to Kant, the ideal of beauty requires a synthesis of understanding and will. Existential philosophy is concerned with conceptual analysis of the most affective aspects of psyche and destroys the popular image of an aloof philosopher. Dostoevsky and Nietzsche can be characterized in many ways, but not as aloof. Shakespeare, although appealing to the affective side of our psyche, at the same time addresses and solves the deepest philosophical questions of the essence of love and meaning of life. Today, the best pieces of art are expected to appeal to both sides of our psyche, thinking and feelings; we expect a synthesis of new differentiated concepts and new differentiated affects. It is a daring task and few are capable of solving it. So, it is not surprising that art goes to extremes, trying to fulfill new expectations. The disproportion between extreme means and shallowness of ideas that we often see today in the world of art is not necessarily due to "degradation" of culture, but is also caused by unrealistic expectations and aims of the artists. And the patient and discriminate lover of art finds contemporary pieces that continue the line of great masters and achieve new levels of synthesis of ever finer differentiated concepts and affects.

Computational intelligence faces an opposite challenge. Creativity of robots is limited by too much differentiation of their psyche. If a robot knows everything it was designed for, it cannot be creative. This is another point of view on one of the main themes of this book: that adaptivity requires fuzziness of the a priori concepts. Aristotelian logic is very limited in adaptivity and creation of new concepts. Fuzzy logic of Aristotelian Forms is needed for maintaining adaptivity, and the mathematical description of this process is provided by MFT. On the one hand, the robotic intellect requires differentiated internal models much wider in scope than current computer capabilities. On the other hand, robotic creativity requires synthesis of the general concepts, beyond the differentiated a priori models; a problem yet to be solved by developing models with evolving hierarchies.

12.2.3.3 Creativity and Unconscious

Human intelligence maintains the source of general, fuzzy, undifferentiated models within its unconscious. Creative expansion of our conscious internal models proceeds by bringing to consciousness the contents of our unconscious. The role of the unconscious in creativity is far from being understood. The creative process can often proceed at unconscious levels; this was repeatedly described by scientists and artists. It seems that there always is an unconscious part of a creative process. Jung concluded that severing connections between consciousness and unconscious leads to a loss of creativity. Our mathematical analysis explains this fact by relating adaptivity to fuzziness of a priori concepts. Although the goal of creativity is differentiated consciousness, completely differentiated consciousness loses adaptivity. It expands the role of Aristotelian logic from the realm of Understanding to the entire psyche and loses the capability for perceiving novelty. (One may find many people like this around, who have a ready answer to every question and are closed to new insights.) Uncertainty and fuzziness are a must for perceiving novelty. And the unconscious is an infinite provider of fuzziness.

Is the problem of robotic creativity going to be solved by developing a psyche complex enough to contain the unconscious?

12.2.4 Mysteries of Physics and Consciousness: New Physical Phenomena?

Several properties of consciousness look mysterious. To me, the mystery is not in those relatively rare parapsychological or other similar events that are difficult to imagine or explain within existing scientific theories. The mystery is that there are basic facts concerning everybody and everyday experience that appear to be impossible to reconcile with rational understanding. Free will is the most important among these facts. Everyone believes that he or she has a free will. The feeling that we can control, at least to some extent, certain parts of our desires and volitions is so fundamental to our existence that we would not give it up, no matter what other considerations might be, including scientific deductions. If science cannot explain free will, most will doubt this aspect of scientific thought, rather than free will. Therefore, science either has to explain free will, or to acknowledge that here is the boundary of applicability of the rational method as far as it exists today.

Freedom implies a degree of unpredictability, but in a way opposite to that of randomness or chaos. Weather patterns are chaotic in that minor variations in initial conditions may cause tremendous consequences some time later; because of this, weather patterns are not computable and not quite predictable. An outcome of a coin toss is random; even a better example of randomness is an outcome of quantum measurement. But none of these examples of unpredictability can explain freedom. The fundamental difficulty is that freedom supposes its opposite, causality. If there is no causality, there could be no freedom. But if the world's laws are causal, how could freedom be explained? Kant's explanation was to consider causality as an a priori concept of mind and freedom as a noumenon, the property of the unknowable human-in-itself.

Some other properties of consciousness are almost as mysterious as free will. These include creativity and the related nature of physical and mathematical intuitions. Possibly, the nature of intuition is the same in the arts and sciences. Scientific truth is similar to beauty, and might be related to unconscious a priori models of psyche that evolved over hundreds of millions of years. But scientific intuition seems "more mysterious" because of an opposition of the fuzzy origin of an intuition to the precision and explicitness of mathematical and physical theories. It would be acceptable to explain by evolutionary mechanisms an intuition about a scientific theory that would give an approximate description of nature, say within one-tenth or even one-hundredth of the actual measured values. But it is very counterintuitive to accept an evolutionary explanation for a physical intuition leading to the development of the general theory of relativity or quantum mechanics, which agrees with measurements up to one-millionth of one-millionth part.

To reconcile these properties of our consciousness with the scientific method, Penrose came to a modified Platonic view that there are three interrelated worlds: the world of consciousness, the world of matter, and the world of ideas, including mathematical objects and constructs. His main point is not to argue with Aristotle, who placed the world of ideas inside our heads, and the three worlds of Penrose may be no different from the Kantian theory of mind or modeling field theory. But Penrose emphasizes that we have to come up with a scientific explanation for all the three worlds. He believes that the new undiscovered yet physical principles of the material world are needed for the description of consciousness. A discovery of these new principles will constitute a theory that he calls Correct Quantum Gravitation. This future theory will unify quantum theory and the general theory of relativity

and will explain the nature of quantum measurement as a nonlocal, nonalgorithmic process. The nonalgorithmic nature of future physics, according to Penrose, will resolve the mysteries of creativity and free will related to the exit out of the finite world of events into the infinite world of ideas.

12.3 Epilogue

The interplay of apriority and adaptivity, transcendence and immanence, eternity and minuteness is a fundamental fact, the condition of human existence. Following the teachings of Böhme and Berdyaev, history is a realization of the idea of an a priori and timeless God in the ever-changing world of matter in space and time. And a single human life is an adaptive realization of the a priori given. And the scientific method, from its origin in medieval scholasticism, to Descartes and Newton, and to today, appears as a mathematical nexus between a priori laws and empirical data—a nexus that in the area of material substance has been accomplished by Newton, and in the area of spiritual substance is being realized today in discovering physical concepts of the mind.

The realistic metaphysics founded by Plato and Aristotle and steadily evolved by thinkers to our time continues today in the physical intuition of spiritual substance as an adaptation of a priori internal models, which combine adaptivity and apriority in that the principles of the models, their functional forms, are genetically inherited and adaptive changes of model parameters and parametric structures in the process of learning provide for a physical premise of an individual cognition.

A priori models, being the main instrument of cognition, at the same time define limits of the accessible to cognition. This dual, antinomial role of a priori models enables us, on the one hand, to understand mind in its everyday manifestation as an adaptation of the a priori model, and on the other hand, to understand creativity as a widening of the a priori aspects of the conscious models—the concepts, imperatives, and postulates of the a priori contents of pure and practical reason. Started by Kant, rational analysis of the concrete contents of a priori knowledge turns out to be a first step toward creative *conscious* enlargement of the postulates of the a priori models. Unconscious creativity, noted even by Plato, is related to a dualism of conscious–unconscious in a priori knowledge, determining two types of creativity—the conscious and unconscious. Describing a creative process of poets and oracles, Plato emphasizes a state of unconscious ecstasy; however, for himself, he accepts a contrary, conscious creative process founded on the rational method of Socrates. The rational aspect of the tradition of Plato and Aristotle, continued by great thinkers for more than 2000 years, points toward an increasing role of the rational principle in the creative process, although not negating the nonconscious roots of the nature of creativity. Thus, the nature of the process of creativity has been changing. The identified direction of this change, when considered from the vantage point of the physics of mind as a modeling field, means that creativity is the differentiation and widening of the boundaries of the conscious parts of the internal models, while maintaining undivided individual psyche through the synthesis. The creative process discovers those parts of the model that are yet hidden in the depths (or hights?) of the area beyond the conscious and integrates them within consciousness.

The spiritual movement of romanticism directly relates to the issue of the nature of creativity (Riasanovsky, 1992). In the theory of mind as a modeling field, romanticism is

understood as an attempt to transcend the boundaries of knowledge, to overcome these boundaries through negation, rejection of a priori knowledge, and not through reason, analyzing the nature of the a priori model. I would like to compare the physical concept of mind founded on the a priori model with the romantic understanding of mind. And I would like to analyze the nature of creativity corresponding to the physical and romantic conceptions of mind. This is of pressing interest because the romantic concept continues to exert a profound effect on artistic, political, and scientific thinking of today, and influences the development of scientific approaches to intelligence.

As a reaction to the antinomial nature of a priori knowledge, there is a romantic dream of a pristine, perfect perception of a precivilized human, unspoiled by education and by concepts that stand between us and the world. Let us trace the influence of this romantic dream in terms of Freud's concepts. Interpreting structures of the psyche such as the superego, which constitute a part of the a priori model, Freud emphasizes its interference with the ideal perception and cognition rather than its role as an organizer of individual experience and a facilitator of perception, cognition, and consciousness. Such a romantic approach does not explain the possibility of perception and cognition of the world, because it does not account for the a priori internal model, which makes perception and cognition possible. As far as we can judge today, an a priori model in animals is simpler and contains fewer differentiable elements than an a priori model in humans. A cognition in a such simpler model results in a world with relatively few definitely differentiable objects, while the rest of the world consists of barely perceived chaotic elementary shapes and motions determined by properties of the retina (something like the perception of a newly appearing objects in human peripheral vision). And even these elementary perceptions are possible to the extent that the retina contains a priori models of these perceptual elements. Simple acts of perception (by senses) or cognition (by mind) are possible because they are the acts of recognition (by senses or mind) of the elements of the a priori model stored in the a priori Forms of the neural organization. There are no trees, nor leaves, nor creeks in the world—only elementary unordered perceptual events—everything organized can be perceived only to the extent that in the neural organization of our brain a priori models are imprinted that complete and organize the elementary perceptions into a coherent picture of the world. The concept of mind based on internal models assumes that the main principles of these models, their functional forms shaped by evolution, are genetically inherited, while adaptive changes of the parameters of the models in the process of learning form a physical basis for individual cognition.

Kant considered that the a priori categories of the mind (the models) constituted the faculty of Understanding. And he described the working of the mind as an interaction among faculties of Understanding, Judgment, and Reason. Judgment, the faculty of feeling, relates the a priori models to objects in the world. And Reason, the faculty of will, directs actions based on the results of Judgment. Actions most important for the intellect are learning by modifying contents of the a priori models. Modeling field theory described in this book provides the mathematical apparatus for the Kantian conception of the mind. It describes mathematically the basic intellectual process of mind as a dynamic symbol, a vortex of input signals, concepts, emotions, and actions.

Modeling field theory provides a foundation for the mathematical theory of the emotional intellect, and for the concept of beauty. Beauty is related to the ability of the faculty of Judgment to perceive pure purposiveness as a potential for the improvement of the internal

models. This aspect of the future physical theory of mind I call cyberaesthetics. It will reveal the perception of beauty as an adaptation property of complex intelligent systems capable of adaptation and learning beyond specific goals. As a property perceived by adaptive systems, beauty cannot be prescribed a priori, but depends on the current state of the internal model. Thus, beauty is not entirely subjective either, it is not at the "whim of a beholder," but is based on objective properties of the subject and subject–object interaction. A possibility for mathematical theory of beauty described in this book is bound to the understanding of the purposiveness of the internal models as it relates to their differentiation and improvement.

How do the a priori models evolve? How do the new Forms-archetypes appear and the bounds of the a priori accessible widen? What is the mathematics of the nature of the creative process and is such mathematics possible? Is the nature of creativity in widening the internal model and what can be said about the physical principles of the creative process? These questions are formulated today as scientific problems as reviewed in Chapter 10, and we are still very far away from a complete understanding. I believe that these issues are related to freedom of will, and, consequently, as shown by Berdyaev, to the meaning of history (Berdyaev, 1969). Penrose believes that new as yet undiscovered physical principles of the material world are needed (Penrose, 1989). The nonalgorithmic nature of future physics, according to Penrose, will resolve the mysteries of creativity and free will related to the exit out of the finite world of events into the infinite world of ideas.

An ancient Chinese proverb says that behind every man there is another man. Everything we build rests on the shoulders of those who built before us. At the end of the book, I would like to express my gratitude to those who's conceptions and discussions influenced my ideas and the content of this book. Since it is impossible to list everyone here, I would only mention our contemporaries, who's ideas are being evolved. Chomsky insisted that language learning is based on a language faculty, an a priori, genetically inherited component of the mind; what I call an a priori model throughout this book. In the 1980s, Chomsky formulated a concept of mind as an abstract system of genetically fixed principles and adaptive parameters. The new concept has been recognized as a radical change in the theory of a priori content of language faculty and as a principal discovery in Chomskyan linguistics (Botha, 1991). But a mathematical realization of the new program cannot be achieved using combinatorial hypothesis testing methods, which are not realizable computationally and do not stimulate the intuition about the exact content and structure of the a priori models. New mathematical methods of combining apriority of the principles-models with adaptivity of parameters are needed, which correspond to the physical intuition of the working of the mind. Modeling field theory provides the needed mathematical apparatus and might help in the discovery of the a priori content of the language faculty.

Dmitriev's conception of the nature of the beautiful in literary texts and his relating the beautiful to the working of mind influenced my development of MFT toward the mathematical theory of emotions. Discussions with Freeman clarified for me inadequacies of classical psychological attitudes to the concepts of mental images and representations as too concrete and nonadaptive.

The concept of the physics of the mind as a neural field of interactions between internal and external signals was originated by Grossberg. His and Carpenter's adaptive resonance theory describes perception as a resonance between internal and external signals. Modeling field theory can be viewed as an extension of this concept toward internal signals being generated by internal models, whose adaptation constitutes learning, leading to an ever

increasing complexity of the internal model. My work on this book tremendously benefited from our discussions. And, I am especially thankful to Steve, who was the first to suggest writing this book.

Holland's mathematical ideas about the nature of the genetic evolution, especially his idea of schemata as an object of evolutionary pressure, enabled me to relate modeling field theory to evolutionary computational concepts. Levine's concepts and discussions helped me to formulate my descriptions of the neural correlates of consciousness. Meystel's conception of the hierarchical semiosis led me toward developing the modeling field theory of the dynamic symbol. Our discussions of the nature of similarity, hierarchies, and semiotics inspired many sections in this book. Minsky's ideas about symbolic artificial intelligence and the society of mind provided significant inspirational impetus, even if often for thinking in the opposite direction. Penrose reminded me very forcefully about the remaining unsolved fundamental problems in the quantum theory. His discussion of Gödel's theory led me to consider its limitations, while discussions of the mystery of the physical intuition inspired my thinking about the development of mathematical foundations of the beautiful.

Fuzzy logic developed by Zadeh turned out to be fundamental to an understanding of the mind and the development of modeling field theory. Without fuzziness of our a priori concepts there is no adaptivity, no learning, and no creativity. Our discussions of adaptivity and granularity inspired multiple passages in this book. Modeling field theory can be viewed as an extension of fuzzy logic in the direction of complex adaptivity and apriority.

How far can we go in building a concept of mind and consciousness founded on an a priori internal model of the world, ascending to Plato's world of ideas, and combining Aristotelian Forms of mind with Zadeh's logic of fuzzy? To what extent will adequate a priori adaptive models emerge in the near future? In the discernment of the concrete content of these models, what will the role of the Kant–Grossberg method, analyzing antinomies of mind be? Looking toward a unified theory of mind and brain, the physics of the mind is being created today in parallel with the physics of the brain, and where would the invisible border separating these two areas be—the border defined by the interplay of a priori and empirical, the physical idealization and physical intuition of mind?

And what is the role of unknown yet physical phenomena in explaining the mystery of the mind?

NOTES

1. "Free will is not a cucumber" responded Nietzsche to those philosophers who were cavalier about explaining the nature of free will. Many cognitive scientists and psychologists today continue this cavalier attitude, whereas others warn about pitfalls of shallow thinking. Concepts of free will and creativity cannot be formulated today as a scientific problem. These concepts contradict causality, and this cannot be reconciled with the science. But, more so, these concepts insist on purpose. Chaos or randomness has nothing to do with freedom and contradict freedom no less than causality. Freedom is deterministic, but noncausal. Noncausal purposiveness cannot be formulated today as a scientific problem. Few scientists have had the guts to acknowledge this; among the brave ones are Descartes and Chomsky. Although Maimonides warned about the importance of understanding where the current limits of rational understanding are. But many scientists prefer a cavalier attitude toward higher mental abilities. A notable example is Minsky's

Society of Mind (1985). A section entitled "Free Will" discusses interactions of multiple agents, none of whom possesses free will. A reader might think that Minsky does not believe that humans possess free will. But in many other places in the book he refers to people as "looking for excuse instead of," etc., as if he believes that humans can alter their destiny by their will. Where is this will coming from?—A thoughtful reader is left bewildered, because the "explanations" of free will do not explain an intuition shared by everyone including Minsky. Many examples of similar discussions of free will are abound (Franklin, 1995). Everyone has an intuition about himself or herself as a person with a free will, but many scientists seems to ignore this intuition when discussing the subject.

2. Nagel thought that subjective phenomena are difficult to explain scientifically because "every subjective phenomenon is essentially connected with a single point of view and it seems inevitable that an objective, physical theory will abandon that point of view." This illustrates a common misperception about the nature of difference between "objective" and "subjective." This could be traced to Kantian "transcendentalization" of the nature of a priori. According to this point of view, a priori transcends *all* experience. But we have seen that this is not necessarily so: a priori may have an adaptive nature. As both the apriority and adaptivity are subjects of science, so both the objective and individual are.
3. The timeless and spaceless nature of archetypes of the collective unconscious is discussed by Jung (1951), who analyzed the primordial origin of archetypes and their uniform nature among various peoples. A relatively recent origin of our a priori intuitions about ordered time and space is confirmed by the observations of psychologists (Lèvy-Bruhl, 1910) and linguists (Whorf, 1936) concerning different conceptions of time and space in different peoples, in particular, some primitive tribes perceive time and space as not quite ordered globally with a higher level of the local orderliness (in the local region of time and space where the tribe currently lives).
4. Notwithstanding, it seems that humans are capable of recognizing (objects) without any specific need. This ability for "disinterested" intentionality is the foundation for our "higher" mental abilities, including an ability to perceive beauty. This ability, however, appears as disinterested only for the purposes of "lower" instincts. Its intentionality is related to the "higher" instinct of acquiring knowledge for its own sake. Clearly, it is related to adaptation, and in the long run to survivability.
5. As far as we know today, the largest part of a large Cretaceous cortex (whales, dolphins) is devoted to acoustical mapping of the environment, including the complicated changing acoustical propagation properties of the underwater world. The Cretaceous brain has almost no interaction between its hemispheres and is more homogeneous than the human brain in other ways as well.
6. Great masters of all times strove for this differentiated synthesis.

BIBLIOGRAPHICAL NOTES

The definition of consciousness is taken from *Webster's Third New International Dictionary*. Merriam-Webster Inc., 1981.

Contemporary scientific understanding of the role of the unconscious is, in a significant part, due to the works of Freud and Jung. Freud discovered the unconscious (1900). Jung developed conceptions of psyche and consciousness based on unconscious archetypes; he identified a number of archetypes and their properties and described psychic types and the differentiation process (1921, 1934, 1951).

On relationships between beauty and truth in literary texts (Dmitriev, 1997).

Historical evidence for the evolution of consciousness from the primordial, to the collective, to the individual is discussed in Jaynes (1976). Primitive consciousness was analyzed by Lèvy-Bruhl

(1910), who introduced the concept of participation mystique. Whorf (1936) analyzed relationships between consciousness and language and found evidence for varying degrees of differentiation of linguistic constructs (grammar).

Recent discussions of neural mechanisms of consciousness and emotions can be found in Grossberg, Levine, Pribram, Taylor, and Zeki. Grossberg (1995) suggested that a necessary condition of consciousness is a resonant state of the type emerging in the ART neural network; the nature of the ART resonance is similar to the MFT convergence state (convergence of thinking vortex process). He further suggested that motor learning is usually unconscious because it occurs via negative feedback error correction, which does not lead to a resonant state. The competitive relational model of Taylor is like a simplified version of MFT; its semantic net and comparison net can be compared to the MFT internal models and similarity measures. Brain mechanisms and areas involved in vision are discussed in Zeki (1996). The limbic system and its functionality was suggested and elaborated by Broca (1878), Papez (1937), and MacLean (1952). Further information can be found in the *Encyclopedia of Neural Science* (Adelman, 1987). One of the early neural network modeling of emotions was Grossberg's gated dipole network (1972a,b), which modeled psychological opponent processes including positive and negative affect, and included emotions as an integral part of learning.

Involvement of the nineteenth-century physicists in neural and psychological investigations is discussed by Grossberg (1988).

Semiotic analysis of psyche is related to semiotic processes in inorganic matter by Taborsky (1998, 1999).

The possibility that neurons perform quantum computations within their microtubular structure was explored by Hameroff (1987, 1994). Penrose (1989, 1994) discusses a need for a new physical theory combining quantum physics and gravity. Quantum theory of modeling fields is described in Perlovsky (1997c).

For Chomsky's theory of language faculty see Chomsky (1972, 1981) and Botha (1991). This is currently a large field of research, with several periodic publications. Some are devoted to studying neural structures associated with language, identifying hereditary deficiencies of specific grammatical concepts, and localizing them in the brain.

LIST OF SYMBOLS

This section contains brief definitions of the main notations used throughout the book.

A_k	amplitude of kth submodel
AI	artificial intelligence
AIC	Akaike information criterion
A-L, A-LL	Aristotelian similarity
AR	autoregressive model or signal
ARMA	autoregressive moving average model or signal
ART	adaptive resonance theory
AZ-LL	fuzzy-adaptive similarity
$[A, B]$	commutator, $AB - BA$
B	Bhattacharyya distance
$\mathbf{C}$	covariance matrix
$\mathbf{C}_h$	covariance of the hth pdf (h submodel)
$\mathbf{C}^{-1}$	inverse of matrix of $\mathbf{C}$
C_{hn}	quantum amplitudes
CAS	complex adaptive systems
CR	Cramer–Rao (theory)
CRB	Cramer–Rao bound
D	dimensionality of a space (or data vector)
$\mathbf{D} = \mathbf{x} - \mathbf{M}$	difference between a vector and its mean
$d(\mathbf{x}, \mathbf{y})$	distance between vectors $\mathbf{x}$ and $\mathbf{y}$
det $\mathbf{C}$	determinant of matrix $\mathbf{C}$
d/dx	derivative with respect to x
$\partial/\partial\mathbf{S}_k$	partial derivative with respect to $\mathbf{S}_k$
$\delta_{hh'}$	delta function
E	entropy
$E\{\cdot\}$	expectation

$E\{\cdot\|h\}$	conditional expectation
$E\{\mathbf{x}\}$	expected value of $\mathbf{x}$, mean
$E\{\mathbf{x}_n\|k\}$	conditional expected value of $\mathbf{x}$, given (class or hypothesis) k
EM	estimation-maximization algorithm
ENN	Einsteinian neural network
ε	photon energy
$f(k\|n)$	fuzzy membership function
$F(\omega)$	(Einsteinian) spectrum model
Φ_ω	number of physical states for a single photon
G, $G(\mathbf{x}\|\mathbf{M}, \mathbf{C})$	Gaussian density
GT	Gödel–Turing theory
GQMF	Gibbs quantum modeling field system
Γ	phase space volume (the total number of quantum states)
$\Gamma_{\mathrm{U},0}$	the total number of states in the universe (for noninteracting IS and W)
$\Gamma_{\mathrm{W}\|X}$	the number of states in W given data X
h, H_k, k	hypothesis (or model, class) number
H	total number of hypothesis (also a set of hypothesis)
H	Hamiltonian
$\mathcal{H}$	Hamiltonian density
HQMF	Hamiltonian quantum modeling field system
$\|h>$	internal QMF quantum states
$\hbar$	Plank constant
I	mutual information
IS	Intelligent system
k	k-factor
k	Boltzmann constant
$\xi(k\|n)$	crisp membership function
Ξ	partition (segmentation, association) of the data among classes (models)
L	likelihood
LL	log likelihood
$\mathrm{L}(\Xi)$	likelihood conditional on segmentation Ξ
$\mathrm{l}(n\|k)$	conditional similarity measure between datum n and model k
$\mathrm{ll}(n\|k)$	logarithm of the conditional similarity measure
$\mathrm{l}(k\|\Xi)$	k-model similarity conditional on segmentation Ξ
$\mathrm{L}(\Xi)$	total similarity conditional on segmentation Ξ
LTM	long-term memory

$\|\lambda >$	eigenstates of $\mathcal{X}$ operator
λ	eigenvalues of this operator
λ	Lagrangian multiplier
m	a submodel index
M	total number of submodels
$\mathbf{M}_k$	kth model
M_{ih}	components of the above vector, $i = 1, \ldots, D$
$\mathbf{M}_h(\mathbf{S}_h, n)$	hth model (mean) of $\mathbf{X}(n)$
$M_{ih}^{:a}$	derivatives, $\partial M_{ih}/\partial S_h^a$
$\hat{\mathbf{M}}$	estimated mean value; "hat" denotes estimated quantities as distinct from the true values of model parameters
$\mathcal{M}$	quantum operator measuring $(\mathbf{M}_h \cdot N_h)$ values
ME	maximum entropy estimation principle
MF	modeling fields
MFT	modeling fields theory
MHT	multiple hypothesis testing algorithm
ML	maximum likelihood estimation principle
MLANS	maximum likelihood adaptive neural system
n	observation number
N	total number of observations (also a set of n values)
N_h	average number of observations classified to hypothesis h (a subset of N corresponding to h-hypothesis)
$n \in k$	(data) element n belonging to class k
NP	number of parameters
$\|n >$	external world quantum states
NRT	nucleus reticularis thalami
$\{N_1 \| \ldots \| N_{\mathrm{H}}\}$	partition of the set $\{N\}$
OC	operating curve
ω	frequency
ω_k	mean frequency of kth submodel
P_{l}	probability of leakage
P_{d}	probability of detection
P_{fa}	probability of false alarm
$P(x)$	probability of an event x
$P(x\|y)$	the conditional probability of x given y
$P(h\|n)$	a posteriori probabilities
pdf	probability density function

$\text{pdf}(x\|y)$	the conditional pdf of x given y
$\Pi_n\ f(n)$	product of $f(n)$ over x
$\Psi(t)$	the quantum state of the universe
QMF	quantum modeling fields
r	correlation coefficient
$r(h)$	a priori probability of the hypothesis h
$\mathbf{R}_k$, $\mathbf{V}_k$	target position and velocity
ρ	density matrix or operator
$\mathbf{S}_h$	parameters of the hth submodel
S_h^a	components of the above vector, $a = 1, \ldots, A$
$S(\omega)$	spectrum
STM	short-term memory
σ	standard deviation
σ_k	kth submodel standard deviation
$\Sigma_x\ f(x)$	sum of $f(x)$ over x
t	time
t_n	time of observation n
T	temperature
Tr	trace of a matrix or operator
U	universe
W	World
$\overline{\mathbf{x}}$	average of $\mathbf{x}$
$\mathbf{X}(n)$	nth observation vector
x, y	variables
$\mathbf{x}$, $\mathbf{C}$	bold indicates vectors and matrixes
$\mathbf{x}^{\mathrm{T}}$, $\mathbf{C}^{\mathrm{T}}$	transposed vector and matrix
$\|\mathbf{x}(n)>$	internal HQMF encoding of the external patterns $\mathbf{X}(n)$
$\mathcal{X}$	an observation operator acting on the internal QMF states
Z-*LL*	Zadeh (fuzzy) similarity
$\langle \cdots \rangle$	average value, usually weighted with probabilities

DEFINITIONS

This section contains brief definitions of the main concepts used throughout the book.

Adalines Artificial neural networks created by Widrow (1959). Used simple internal models.

Adaptive fuzzy membership A *fuzzy membership*; the fuzziness can adapt to the data.

Adaptive fuzzy similarity (AZ-similarity) A similarity measure between the model and the world that combines adaptive segmentation and low computational complexity.

Adaptive resonance theory (ART) A neural network developed by Carpenter and Grossberg (1987), which describes perception as a resonance between afferent and efferent signals, that is between signals coming from the outside, from sensory cells receiving external stimuli, and those coming from the inside, that is signals generated by a priori models. ART is a theoretical principle of the structure of adaptive robust feedback connections between two different levels of a neural network. One level is cognitively "higher" than the other.

Affect An unconscious undifferentiated emotion that is not under the control of free will.

Adaptivity An ability of the mind to adapt to changing environments, to learn from experience.

AI An abbreviation for artificial intelligence.

A priori A priori knowledge refers to the content of the mind prior to experience. Classical philosophers considered a priori as transcendent and prior to any experience whatsoever. This view of a priori is still used by many philosophers today, however, it is becoming less and less useful in the face of increasing appreciation of the development of the a priori contents of the mind: genetic evolution, embryo development, early childhood development, continuous learning, and adaptation. Therefore, throughout this book, I take a more pragmatic and more theoretically palatable view of what is a priori: it depends on the context. When appropriate, a priori is understood as including all the contents of the mind prior to the most current experience.

Apriority An ability of the mind to utilize a priori knowledge (available before experience, genetic, inborn, God given, learned previously).

Apriority vs. adaptivity of mind One of the main philosophical issues debated through millennia under various names, including materialism vs. idealism, realism vs. nominalism, immanence vs. transcendence, quidditive vs. existential aspects, behaviorism vs.

mentalism, internal representations vs. Pavlovian reflexes, parametric vs. nonparametric estimation, afferent vs. efferent signals, parallel vs. serial processing, connectionist vs. symbolic, learning vs. programming, emergence vs. analytic descriptions, neural vs. symbolic, top-down vs. bottom-up processing, connectivism vs. logic.

Archetypes Primordial structures of psyche discovered by Jung. He assumed their uniform nature among various peoples and calls them the *collective unconscious.* Conceptually related to *Aristotelian Forms.* In modeling field theory they are mathematically represented as *a priori models.*

Aristotelian Forms The Aristotelian theory of mind is based on Forms. Forms are different from Plato's Ideas in that they are dynamic entities that support learning. Forms combine *apriority* and *adaptivity*. A priori Forms exist as potentialities. They become concepts in the process of interacting with the material world. This process constitutes learning.

Aristotelian logic Logic is a science of correct reasoning and of criteria of validity of thoughts. Aristotelian logic addresses absolute eternal truths. Its cornerstone is a "law of excluded third": every proposition is either true or not true (anything else is excluded). It is a foundation for most of mathematics and most of our algorithms of intelligence. Aristotelian logic addresses Platonian Ideas rather than Aristotelian Forms. Computational difficulties of the algorithms based on Aristotelian logic can be traced to the original contradiction in Aristotelian theories.

Aristotle vs. Plato Philosophers emphasized the *ontological* difference between their teachings, which is unimportant today. The difference in *epistemology* of Plato and Aristotle, which is crucial for the design of learning systems, escaped philosophical scrutiny.

Artificial intelligence (AI) In a wide sense, the entire area of mathematical or computational intelligence. In a narrow sense, a specific mathematical approach that uses systems of logical rules; I also call it the Plato–Minsky approach (often called "symbolic AI"; however, I do not use this term because it presupposes a too simple nature of *symbol*).

Assignment A particular type of *associations*.

Association An important step in many intelligent algorithms that consists in establishing relationships between subsets of data and classes (submodels, agents, objects, tracks, processes, sources of signals, types, modes, etc.). Also called segmentation, partition, clustering, grouping, classification.

Autoregression *Regression* applied to time series.

Backpropagation A neural network concept combining a *nonparametric* structure of a classification boundary and gradient learning. A mathematical realization of the *nominalistic* concept of intellect by remembering from past experiences a boundary separating classes or concepts in the *classification space.* Faces *combinatorial complexity* as a *nonparametric technique*.

Bayesian decision theory A mathematical approach to solving the *hypothesis choice* problem based on the theory of probability. Developed by Bayes in 1763. It was the first mathematical technique to combine a priori knowledge with data in the face of uncertainty. A set of hypotheses represents the a priori knowledge. A decision is

made a posteriori, that is after the current data become available. Bayesian theory does not explain learning; still it represents one aspect of *Aristotelian theory of Form*: meeting between the a priori Forms (hypotheses) and matter (data). Under certain conditions, Bayesian theory leads to optimal decisions (for example, for making bets in card games).

Bayesian similarity or likelihood Measures a degree of stochastic deviation of data from its mean value given by the model

Beautiful Beauty is a perceived *purposiveness* (of our *internal representations* in their relationships to the outside world) as divorced from any specific lower level "utilitarian" goal. In MFT, similarity measure is an ability of this type. The object is called beautiful to the extent that its purposiveness is felt in its pure form and is bound to its a priori nature. The nature of beauty is related to an interest not in the object, but in the subject: what I make out of this representation in myself. Beautiful is what coincides with the purpose of acquiring more knowledge and improving the harmony between the internal model and Nature. Beauty is a mechanism of evolution, adaptation, and survival. As with any other survival mechanism, there are mechanisms of camouflage, counterfeiting, and countercounterfeiting. Beautiful is perceived through the aesthetical aspect of *Judgment*. It is related to the "pure" purposiveness of our representations, which is separate from any specific purpose for which an object can be used, and includes only the knowledge itself. MFT provides a foundation for the mathematical description of the beautiful: similarity measures establish emotional relationships among data and models, and activate actions of adaptation toward improving the harmony between the models and nature.

Behaviorism A scientific direction and an accompanying intellectual and philosophical movement, defining psychology as a science of human behavior. Behaviorism attempted to explain the entire human psychology as a sequence of stimuli and reflexes and denied a need for consciousness in understanding of the intellect. It dominated American psychology from about 1920 to 1960. Behaviorism created a scientific methodology of experimental psychology; however, as a philosophy maintaining that the concepts of *consciousness, free will,* and *idea* are not needed in psychology and should be discarded, behaviorism exerted an inhibiting influence on the development of concepts of mind.

Bottom-up processing A mathematical technique of deriving classes and concepts (top level) from the data (bottom level). Related to the principle of *adaptivity* and the philosophy of *nominalism*.

Classification Engineering application areas and mathematical algorithms that classify patterns in data as belonging to particular classes or types.

Classification space A multidimensional space in which events or objects are represented as points (vectors).

Classifier A boundary between decision regions in *classification spaces* (same as *discrimination surface*).

Collective consciousness Most of our conscious models are collective in nature in that they are conditioned by culture and only their relatively minor aspects are adaptive objects of the free will. Collective consciousness that emphasizes individual values

should not be confused with *individual consciousness*, which is a historically new and still rare phenomenon among humans.

Collective unconscious Primordial psychic structures, *archetypes*, discovered by Jung. The contents of archetypes are not accessible directly to the consciousness. They provide a framework, a possibility for the psyche. Archetypes are conceptually related to *Aristotelian a priori Forms*, and to the primary matter of Avicenna. Archetypes initiate adaptive processes in the psyche that may become conscious, but the word "archetype" refers to the unconscious nonadaptive primordial content. Archetypes are mathematically described in MFT by a priori models.

Combinatorial complexity A ubiquitous problem of the algorithms of intelligence. On the one hand, intelligence should be flexible enough to manipulate various combinations of multiple elementary notions, concepts, and plans in order to find a suitable one in complex situations. On the other hand, evaluation or learning of combinations leads to a combinatorial explosion: the number of combinations, even for problems of moderate complexity, exceeding the number of particles in the universe. Algorithms associated with *apriority* faced logical complexity, and those associated with *adaptivity* faced training complexity. Attempts to combine the two led to combinatorial complexity of computations. Combinatorial complexity is also known as the "curse of dimensionality" (Bellman). It was traced in this book to Aristotelian logic and resolved using adaptive model-based fuzzy logic.

Complex adaptive systems (CAS) Systems of intelligent agents that can aggregate and evolve according to genetic algorithms. Proposed by Holland (1992, 1995).

Connectivism A point of view that neural type architectures, composed of a large number of interconnected simple elements (neurons), are essential for explaining mind.

Consciousness An awareness or perception of inward psychological facts, a subjective experience of sensing, feelings, or thoughts. Consciousness directs the will and results in a better adaptation. In complex situations, various instincts might encounter contradictions. Consciousness can resolve an instinctual impasse by suppressing some processes and allocating power to others. By differentiating alternatives, consciousness can direct a psychological function to a goal. Consciousness is a complicated differentiated phenomenon that is characterized (very roughly) by the following levels: undifferentiated awareness, collective consciousness, and individual consciousness. Both collective and individual consciousness are characterized by multiple modalities or types of consciousness that at the top level include sensing, feeling, thinking, and intuition. Conscious parts of *internal models* are more differentiated, more adaptive, more accessible to will and *Reason*, and more amenable to future differentiation and adaptation.

Contradiction law A law of Aristotelian logic, according to which every statement (concept) is either true or false. It is also called the law of excluded third.

Cooperative and competitive dynamics Types of neural organization and interactions among neurons or agents. Cooperative neurons enhance each other's activation levels; competitive neurons inhibit each other.

Cramer–Rao bounds (CRB) Fundamental mathematical bounds on learning, establishing the minimal learning requirements in terms of the amount of data needed to learn the model parameters as a function of the true model structure and parameters.

Creativity An ability to create what did not exist previously (within the collective or one's own psyche). Characterized by multiple levels: lower levels include adaptivity and learning based on a priori models; higher levels include expanding the a priori model, and perceiving the beautiful and the sublime. In this book I use creativity only with respect to the higher levels. Creativity is related to fuzziness of existing a priori models, because without fuzziness there is no learning. It is related to the unconscious, which is an eternal source of fuzziness. It is related to individual consciousness, for individuality requires creation of individual models out of collective ones. The nature of creativity changes throughout history toward an increased conscious element. And creativity possesses a mysterious flavor due to its relationship to *free will.*

Crisp concept A concept of Aristotelian logic. It either matches the data or does not [the third possibility of a partial match is excluded by the law of contradiction ("excluded third")].

Cyberaesthetics A future science of intellectual emotions. An aspect of the physical theory of mind, which will provide a mathematical description of higher emotions, including the beautiful and the sublime. It is based on Kant–MFT theory that describes mathematically the basic intellectual process of the mind as a dynamic symbol, a vortex of input signals, concepts, emotions, and actions. And it will reveal the perception of beauty as a property of complex intelligent systems capable of adaptation and learning beyond specific goals.

Designatum The object to which the sign refers. Its mathematical description in MFT is given by the incoming data associated with the sign.

Differentiation A process of evolution of consciousness, which proceeds through differentiation of psychic functions and their contents. It was introduced by Jung. The original archaic state of consciousness is an undifferentiated identity. And only gradually, psychic functions differentiate (thinking from feelings, from sensing, etc.) and the faculty of Understanding acquires a large number of highly differentiated internal models-concepts, which make differentiated thinking possible. The MFT techniques for estimating and increasing the numbers of submodels as needed describe mathematically an aspect of this process.

Discriminating surfaces A concept according to which learning is a search for a collection of planar surfaces making up a boundary separating classes in the *classification space.*

Ego The specific internal model responsible for consciousness. The main properties of Ego are that it is a subject of free will and it contains a conscious part of Self (the archetype, or internal model), which, to a significant extent, coincides with consciousness. Individuality as a total character distinguishing an individual from others is another characteristic of Ego. Not all aspects of individuality are conscious. Even though the Ego is a product of culture and upbringing to a greater extent than it is a product of personal free choice, free will exists within the conscious part of the Ego.

Einsteinian likelihood A *likelihood* that considers the unknown state of the world as the main source of uncertainty. Usual statistical likelihood considers the main source of uncertainty to be random deviations of the data from the model, as, e.g., measurement errors.

Einsteinian neural network (ENN) A neural network implementing *modeling field theory* based on *Shannon–Einsteinian similarity* (*Einsteinian likelihood*, or information).

Emotions Feelings and their manifestations. An a priori faculty of psyche, an ability to perceive satisfaction or dissatisfaction of basic instincts. Emotions originate from evolutionary-old affective brain systems that control behavior essential for survival and reproduction. Therefore emotional control of consciousness and behavior is common. Unconscious undifferentiated emotions that are not under the control of free will are called affects. Even very strong emotions could be highly differentiated, complicated, completely conscious, and under the control of free will. The conscious control of emotions seems to be of a relatively recent origin and is less understood. Jung considered emotional consciousness. According to our analysis, it is based on structured similarity measures, and its differentiated development is based on the internal models of emotional concepts of relationships and the archetypes of "others." I differentiate between lower emotions related to ancient affective systems necessary for survival and higher intellectual emotions related to Kantian Judgment. A mathematical description of higher emotions in modeling field theory is given by similarity measures. Higher emotions are related to an ability to perceive beauty.

Epistemology The study of the origin of knowledge (including mechanisms of learning). According to Plato, knowledge exists in its final form in the world of Ideas, and our knowledge is due to a mysterious connection to this world. According to Aristotle, knowledge dynamically emerges in the process of meeting between the Forms and matter. The epistemological problem of science has not been solved by philosophy. It is impossible to accept any of the existing theories of the growth of scientific knowledge, such as the received instant rationality of falsificationism, Kuhn's irrationalism, or Lakatos' rationality of continuous growth (see, e.g., Lakatos and Musgrave, 1970), for none of these theories addresses the fundamental issue of the relationship between the growth of science and the content of a priori knowledge.

Evolutionary computation Mathematical techniques of learning that resemble genetic evolution.

Excluded third law A law of Aristotelian logic, according to which every statement (concept) is either true or false (any third alternative is excluded). It is also called the law of contradiction.

Factor analysis A mathematical technique for the analysis of statistical correlations in multidimensional spaces. It models stochastic or random deviations about the mean value, with the mean defined by a single multidimensional deterministic phenomenon. Developed by Spearman (1910) and Thurstone (1947).

Feelings Internal manifestations of a psychic ability to perceive satisfaction or dissatisfaction of basic instincts. In MFT they are described by measures of similarity. See *emotions*.

Formalism A direction in mathematics that formally defined mathematical objects in terms of axioms or rules. "Being unable to intuit or know the objects of science in themselves, we must settle for the formal laws they satisfy" (Webb, 1980). Related to Aristotelian logic. Founded by Hilbert. Inconsistency of formalism was proved by Gödel.

Free will According to Webster's dictionary, the ability to choose between alternative possibilities in such a way that the choice and action are to some extent creatively determined by the conscious subject. A subject of free will is Ego (an internal model responsible for consciousness). Free will is inside consciousness. Free will is limited by laws of nature in the outer world and in the inner world by the unconscious aspects of Self. Free will belongs to consciousness, but not to the conscious and unconscious totality of the psyche. Free will is a mysterious feeling in that it has no rational explanation: free will is opposite to determinism, but it is also opposite to randomness or chaos. Free will is related to the sense of our destiny, and currently seems beyond scientific understanding.

Fuzzy concept A concept of fuzzy logic. It can match data partially and thus can be used as a foundation for learning.

Fuzzy logic Logic is a science of correct reasoning and of criteria of validity of thoughts. Fuzzy logic created by Zadeh addresses relative truths of everyday life. In this book, it serves as a foundation for the modeling field theory, which extends fuzzy logic to combining adaptivity with apriority. Fuzzy logic addresses Aristotelian Forms. Computational difficulties of the algorithms based on Aristotelian logic are overcome using fuzzy logic.

Fuzzy membership A degree of a pixel (or data subset) belonging to a class.

Genetic algorithms Mathematical techniques of learning that resemble genetic evolution. Designed to combat the combinatorial explosion.

Gödel theorems Gödel proved that formal systems related to Aristotelian logic or logic of predicates are fundamentally limited. Turing has reformulated this result for computational systems and demonstrated limitations of any system of algorithms. There have been several attempts to use these results for proving the principled difference between the mind and machine. A most recent one is due to Penrose, who believes that the Gödel–Turing limitations have to be surpassed in order to model the mind. My conclusion in Chapter 11 is that Gödel–Turing results establish limitations to Aristotelian logic, and are related to the combinatorial explosion of complexity of intelligent algorithms, but are not necessarily relevant to the theory of mind.

Gödel theory Established fundamental limitations of logic.

Gradient learning A computational concept of learning in which performance of an algorithm is gradually improved by modifying parameters of the algorithm along the gradient of the performance measure. Also called "hill climbing." It can be used in *model-based* and *nonparametric* techniques, such as backpropagation neural networks.

Heterarchical architecture An organization of an intelligent system that combines multiple modules that have significant independence with a "soft" hierarchy within certain modules.

Hierarchical architecture An organization of a multilevel intelligent system, with each level processing data received from a lower level and reporting results to a higher level. Strict hierarchies have no feedbacks or vertical loops (among levels). According to psychological and neural data, the mind and the brain are not strict hierarchies, but combine multiple modules that have significant independence with a "soft" hierarchy within certain modules.

Hypothesis choice A classical problem of mathematical intelligence: a decision must be made based on available data. A decision consists in selecting one of several available hypotheses concerning what the data might tell.

Ideas When capitalized refer to *Platonian Ideas*, which exist in their own world.

Individual consciousness Conscious aspects of internal models are more adaptive than unconscious ones and they are more affected by personal experience, upbringing, and culture. In particular, consciousness leads to the development of the individual personality, that is, to the development and adaptation of internal models that differentiate persons from their environment in many different ways, while preserving personal identity by a synthesis of differentiated models into a coherent conscious whole. Jung called this process individuation and considered it to be the most important task of personal spiritual development. Most of our conscious models are collective in nature in that they are conditioned by culture and only their relatively minor aspects are adaptive objects of free will. Individual consciousness is the ability to modify the structure of Ego by expanding the conscious parts of the internal models. It is related to free will. All of us believe that we have individual consciousness. But it should not be confused with individualistic *collective consciousness*. Individual consciousness is a historically new and still rare phenomenon among humans.

Individual unconscious Unnoticed consciously, suppressed, or forgotten contents of the individual experience.

Information (Shannon's) A measure of certainty about choice among alternatives. Computation of the numbers of alternative states depends on the goal: which variations are of interest or importance and should be counted as different states, and which should be ignored? Such a general formulation could lead to a definition of information contingent on meaning of various states and on defining intelligence. In Shannon's theory, an engineering problem is limited to counting the numbers of predefined states, such as letters in the alphabet. Thus, Shannon's information is not related to an emergent meaning in a learning system.

Instinct of world modeling Maximizing the similarity between the internal model and the outer world data. A good correspondence between the internal models and the world is so important that there have to be very basic biological mechanisms driving toward regular or even constant improvements of the model. Many forms of exploratory behavior can be explained by assuming a basic instinct or drive to improve the internal model. The MFT similarity maximization mechanism is a possible mathematical description of that instinct or drive to improve the world model.

Intelligence A clear definition of intelligence would appear when its theory comes to near completion. As a first step, let us characterize intelligence as a goal-directed functioning in artificial and natural systems. This includes some of the following: functioning inside and outside of an intelligent system self; selection of goals and subgoals; sensing, perception, recognition, decision, planning, and acting; acting inside and outside of self; learning and adaptation; memory; acquiring, storing, and using knowledge; internal representations; hierarchical and parallel organization (of all of the above: goals, functioning, knowledge); reproduction; evolution; social organization; organization of the environment; and organization of Self. This list is continued toward thinking, feeling, emotion, intuition, consciousness, free will, and creativity.

This book develops the idea that there are specific elements of intelligence. They involve a dynamic process of concept formation in which fuzzy a priori concepts interact with input signals to form new concepts, which are crisp or less fuzzy than the a priori ones. This process of concept formation employs a mechanism of interaction between concepts and emotions. A mathematical description of this process is developed throughout the book (mostly in Chapter 4). This mechanism is uniformly employed at multiple levels of a heterohierarchical organization: in perception (formation of percepts from sensory signals) and in cognition (formation of new concepts from previously learned concepts). The heterohierarchical organization as well as a variety of a priori concepts and experiences determine the richness of the intelligence. An appreciation of these mechanisms as fundamental elements of intelligence is helped by relating the mathematical concepts to the concepts in philosophy, semiotics, and psychology. This relationship is interspersed throughout the book, mostly in Chapters 3 and 10.

Intelligent agent A psychic process or a software module in an intelligent system. Characterized by a degree of independence from other agents in terms of its goals, procedures, initiation, and termination conditions. An MFT agent continuously exercises a sequence of the three Kantian faculties: computes a submodel (Understanding), evaluates a similarity measure and fuzzy membership (Judgment), and acts by learning (changing model parameters) or by sending behavior signals to other agents (Reason). An MFT agent is a dynamic symbol, a vortex of thinking–feeling–action.

Intentionality A property of referring to something else, *purposiveness*. The models and similarities of MFT are constructed so that they have an intent, purpose, or meaning within the intelligent system, which is the mathematical description of the intentionality of the intellect. This intentionality includes the correspondence to the world and adaptivity that provides for learning. Intentionality provides a background for a mathematical theory of higher faculties of mind, including the possibility for mathematical treatment of the beautiful and the sublime. I disagree with recent attempts to relate intentionality exclusively to consciousness: within a living system everything is intentional (directed at survival, reproduction etc.) Individual creative consciousness is capable of abstracting from this lower level intentionality and is intentional or purposive without any specific lower level utilitarian goal; its intent is self-realization, or the future self. The intent of the individual consciousness is the essence of our existence, which is today beyond scientific analysis. (Compare to *purposiveness*. I think that Kant's *purposiveness* is a more accurate term than the now popular *intentionality).*

Internal model Mathematical models used by model-based algorithms. Models anticipate (predict or model) the input data. Models can combine adaptive and a priori aspects of the mind. The internal model is among the most important concepts of mathematical intelligence. Mathematical internal models are related to Platonian Ideas, Aristotelian Forms, Kantian categories of pure reason, and Jungian archetypes.

Internal representations A notion that mind utilizes internal representations of the objects in the process of recognition (similar to internal models, but historically, this notion was overinterpreted: many researchers hold implied specific assumptions about the nature and properties of the representations, and these assumptions could differ widely).

Interpretant A signal indicating a recognized concept inside the mind or an intelligent system. The result of Judgment in Kantian theory. In MFT it is an output a_k from the k-th model reaching a high similarity value.

Judgment In Kantian theory, it is an ability to see that a particular case comes under the general rule. Described in *Critique of Judgment* (1790). It corresponds to the feeling mode of consciousness. In modeling field theory, it is described mathematically by *similarity measures*.

Kalman filter A target track algorithm that combines the track model (of any complexity) with uncertainty due to measurement errors. It was originally developed for a single target (process) and does not perform association.

Kant–Grossberg method Kant explored antinomies of reason to elucidate the explicit a priori contents of mind. Similarly, Grossberg explores visual illusions to elucidate the explicit a priori contents of the visual system.

Kant–MFT theory A combination of Kant's philosophical theory of mind with the mathematical apparatus of MFT. A dynamic system in which the three abilities identified by Kant exist in the process of constant interaction, as if it were in a "vortex." This vortex models learning of a concept as a dynamic formation of a symbol. It provides a foundation for a physical theory of mind, including the concept of beauty.

Kant theory of mind A rational philosophical theory in which mind consists of the three main a priori faculties: Understanding (concepts), Judgment (that a particular piece of data corresponds to a particular concept), and Reason (behavior). Kant goes into great detail, specifying a number of finely differentiated properties of mind and their interrelationships, as if writing a system specification document for a software project. Kant gives a rational explanation of the concepts of the beautiful and the sublime.

Language faculty The a priori contents of mind, which, according to Chomsky, determine our ability to learn and use language.

Learning Same as *adaptivity*. Some researchers differentiate learning as a higher aspect of adaptivity, which is accompanied by structural changes within the learning system. Learning, as adaptivity, requires *fuzzy logic*: a crisp (Aristotelian) concept either matches data or does not, and therefore is incapable of perceiving a need for adaptation (due to imperfect match).

Likelihood *Probability density function (pdf)* considered as a function of model parameters and fixed data.

Limits on learning Fundamental mathematical limits depend on the amount of available data and contents of the a priori models.

Limits to scientific method An important part of this investigation is a delineation of what we can hope to understand from a rational scientific point of view, and what is currently beyond such hope. The line delineating boundaries of applicability of the scientific method is a moving one; still, it needs to be identified so that our scientific discussions could be properly focused.

Likelihood ratio test A method of hypothesis choice according to the likelihood (or pdf) ratio.

Logic A science of correct reasoning and of criteria of validity of thoughts. This book considers *Aristotelian* and *fuzzy logics*. The word "logic," when used without a qualifier in this book, designates *Aristotelian logic.*

Mathematical concepts of intelligence There are a few basic "classical" computational concepts underlying most of algorithms of intelligence and neural networks. These concepts are directly related to the philosophical discussions of apriority vs. adaptivity. Each concept faces a combinatorial explosion. New emerging computational concepts addressing the combinatorial complexity include complex adaptive systems and genetic algorithms, hierarchical organization, and modeling field theory.

Maximum information (MI) A method of estimating model parameters from data by maximizing the mutual information between the data and model. Suitable for approximate models.

Maximum likelihood (ML) A method of estimating model parameters from data by maximizing the likelihood function. The ML has important theoretical advantages, when the model is accurate.

Maximum likelihood adaptive neural system (MLANS) A neural network implementing modeling field theory based on Bayesian similarity (likelihood).

Meaning The meaning of a concept is determined by its interrelationships with other concepts and by actions initiated (within an intelligent system or in the outer world, by activation of the concept).

Mentalism Maintains that complicated mental processes are essential for understanding human behavior and mind. Mentalism opposes behaviorism and is accepted by cognitive science.

Model-based neural networks Neural networks whose structure and learning mechanisms are determined by internal models. In addition to the model-based neural networks considered in this book, other types of neural networks incorporate statistical mixture models similar to MLANS: SPNN (Streit and Luginbuhl, 1990, 1994); HME (Jacobs et al., 1991; Jordan and Jacobs, 1994) and HMD and POEM (Kumar and Manolakos, 1996; Baggenstoss, 1997). A closely related probabilistic neural network (PNN) (Specht, 1990) estimates pdf using a nearest neighbor type Parzen estimation.

Model-based recognition Mathematical algorithms that recognize patterns (events, objects, images, etc.) in input data by utilizing mathematical models of these patterns.

Model-based vision (MBV) Same as *model-based recognition* as applicable to image data.

Modeling field theory (MFT) A mathematical theory combining a priori knowledge with learning and fuzzy logic as a potential approach to physically acceptable concepts of intellect. It is a set of dynamic equations maximizing AZ-similarity. MFT provides a mathematical description of *Aristotelian Forms, Kantian theory of mind,* and the process of *semiosis*, or dynamic process of *symbol* formation. MFT combines *apriority* and *adaptivity* and resolves the conundrum of *combinatorial complexity.* It consists of the three a priori ingredients or faculties: *internal models* that ascend to Plato and Aristotle, measures of similarity between the internal models and the input data that

ascend to Kant's *Judgment*, and the dynamic laws of adaptivity that maximize the similarity between the models and data (ascending to Kant's *Reason).* MFT is a dynamic system in which the three abilities identified by Kant exist in the process of constant interaction, as if it were in a "vortex." This vortex models learning of a concept as a dynamic formation of a *symbol*.

Modeling field theory of consciousness A physical theory of consciousness that can only be outlined at present. In this theory, consciousness is due to an internal model.

Multiple hypothesis testing (MHT) A mathematical method of concurrent *association* and parameter *estimation*. Combines apriority and adaptivity. Suffers from *combinatorial complexity*.

Multiple hypothesis tracking *MHT* method, when applied to *tracking*.

Nearest neighbor concept A nonparametric computational concept of learning: new events or objects are classified to the same class or category as the nearest (most alike) event from past experience. It is the simplest mathematical realization of the nominalistic concept of intellect, according to which *ideas* and *concepts* emerge in the process of learning from experience as names of classes of similar objects (and not from a priori knowledge). It is a straightforward and highly intuitive concept and it serves as the basis for a large number of algorithms and neural networks. It leads to combinatorial explosion of the number of past experiences required for learning.

Neural fields An intuition about and mathematical techniques of the mind as a distributed dynamic (spatiotemporal) process.

Neural networks Biological: an interconnected network of neural cells (neurons). Artificial: devices or algorithms that resemble biological neurons architecturally, functionally, or conceptually.

Nominalism Philosophy created by Antisthenes, founder of the Cynic school of philosophy. Nominalism considers ideas to be just names (nomina) for classes or collections of similar empirical facts. Among most prominent nominalists is Occam, who lived in the fourteenth century. Despite the fact that most of the great scientists were realists (Newton, Einstein), nominalism plays an important role in the scientific method, because it emphasizes the experiential origin of knowledge.

Nonparametric techniques Mathematical techniques of intelligence and learning in which an algorithm uses a large number of parameters that are not directly related to an a priori model of the data or process. Faces combinatorial complexity of training requirements.

Ontology A philosophy of the being. (In this book it usually refers to Platonian ontology of Ideas, as true or more real beings than objects of everyday experience.)

Parametric structures Parametric models, the structure of which is a parametric function. A simple example is the number of active agents in MFT. A complicated example is schemata in CAS.

Parametric techniques Mathematical techniques of intelligence and learning, in which an algorithm uses adaptive parametric models of the data or process.

Partial similarities Similarities between a pixel and a model.

Pattern recognition Algorithms of intelligence directed at recognition of patterns in data. Many pattern recognition algorithms are based on statistical techniques. They use classification features and many are adaptive.

Perceptrons Artificial neural networks created by Rosenblatt (1958). Used no prior knowledge.

Personality types Determined by a predominant mode and attitude of consciousness. Jung identified four modes of consciousness: thinking, sensing, feeling, and intuition and two attitudes: introverted and extroverted.

Plato–Minsky approach A specific mathematical approach to computational intelligence that uses systems of logical rules. Related to the principle of *apriority* and the philosophy of *realism*.

Platonian Ideas A concept of mind that maintains that our ability to think is founded on the principle that concepts or abstract ideas of mind are known to us *a priori*, through a mystic connection with the world of Ideas. Ideas exist in their own world.

Philosophical concepts of intelligence Much of the philosophical discussion of intelligence has being devoted to the "mind–body problem." It was concentrated, to a significant degree, on the roles of *apriority vs. adaptivity*. Another area of the discussion was shaped by Kant and concentrated on the explicit contents of a priori knowledge. Most recently, debates are about emotions, the beautiful, subjective experience, consciousness, free will, and creativity.

Physical theory of mind Combines physical intuition (originated in a priori contents of our mind) with mathematics and experience. Eventually it combines the mind and the brain into a unified theory of spiritual and material substance. Currently it is being developed along several lines. For example, Grossberg combines the intuition with data that are primarily psychological and neural. I combine the intuition with data that are primarily the philosophical concepts of mind.

Pixel similarity Similarity between a pixel and all submodels.

Pragmatics In semiotics: relations between sign-vehicles and their interpreters. Its mathematical description in MFT is given by the adaptation actions of Reason, modifying the models, and other actions, such as sending a message to other agents.

Probabilistic data association (PDA) A tracking algorithm in which all measurements are probabilistically associated with each track using a posteriori Bayes probabilities, followed by the parameter update computed using a Kalman filter. It avoids combinatorial complexity but requires good track estimates already formed. Otherwise suffers from combinatorial complexity.

Psychic functions, or modes of consciousness Jung identified four main functions: thinking, feeling (emotion), intuition, and sensing. Thinking and feeling are rational and more easily accessible to consciousness and more easily attain highly differentiated status. Intuition and sensing are irrational and less accessible to consciousness and differentiation.

Psychology of philosophy The philosophical schism between *realism* and *nominalism*; according to Jung (1934) it has been, to a significant degree, due to the antagonism between two different psychological types: the introverted and extroverted. Thinkers

of the introverted type are more conscious about their internal thoughts and tend to emphasize a priori internal knowledge, whereas thinkers of the extroverted type tend to emphasize learning from experience.

Pure Reason A Kantian term designating the world of general concepts as a specific content of our mind.

Pure Spirit Designates a pre-Kantian, precritical philosophy of mind.

Purposiveness From the point of view of an intelligent system, like ourselves or a robot, every object, as a phenomenon, has the purpose of being recognized by an intelligent system (in addition to any other purpose to which we or the robot may put this object). The universal purpose of any object is its concept: for an object to have any purpose for a particular intelligent system, the object's concept has to exist in the system. This is a design principle of any intelligent system. And this design principle is applicable to us: evolution (or God) designed us so that we can find our way around those objects that we recognize in nature. The basic principle of the design is that nature appears to us as purposive. The purposiveness of nature is the a priori part of our representations and it harmonizes nature with our desire for knowledge and produces the feeling of pleasure (or pain, if chaos is encountered). (Compare to *intentionality*. I think that Kant's *purposiveness* is a more accurate term than the now popular *intentionality*.)

Pyramid of self-reflections Within the hierarchical internal model, on every level, the model of self contains a submodel of the previous level model-of-self, resulting in a pyramid of self-reflections (Meystel, 1995).

Qualia Subjective feelings accompanying perception. Some contemporary philosophers suggested that the subjectivity of qualia is a mystery beyond mathematical description. I disagree. Qualia do not represent a mystery for the concept of mind based on the internal model. Within a context of two-way interaction between individual sensations and the a priori model, the subjectivity of qualia is explained by the adaptivity of the model, which includes inherited structures, individual development, and the remains of the entire accumulated experience up to the moment of a particular individual perception. In addition, one has to account for conscious as well as unconscious aspects of the internal model. This physical intuition about qualia presents no mystery to mathematical description.

Quantum measurement A process of relating a state of a quantum system to the macroscopic classical world accessible to our conscious perception. A quantum system is described by a wavefunction, which is a superposition of multiple states. During a "macroscopic observation," these multiple states "collapse" to a single macroscopic classical state. Existing quantum theory provides us with a mathematical technique leading to an extremely accurate description of the observed physical world, but the process of the wavefunction collapse remains unexplained. Penrose (1989) believes that a future theory that he calls Correct Quantum Gravitation will unify the quantum theory and the general theory of relativity and will explain the nature of quantum measurement as a nonlocal, nonalgorithmic process. The nonalgorithmic nature of future physics will resolve the mysteries of creativity and free will related to the exit out of the finite world of events into the infinite world of ideas.

Realism Philosophy created by the school of Plato and Aristotle. According to Plato, Eide or Ideas *really* exist in their own world. Our ability to think is founded on the principle that concepts or abstract ideas of mind are known to us *a priori*, through a mystic connection with the world of *Ideas*. Realism postulates the *apriority* of mind.

Reason In Kantian theory it is an ability to draw conclusions that will generate behavior. (The most important type of behavior, interwoven with higher intellectual abilities and emotions, was considered by Kant to be the behavior of learning.) Described in the *Critique of Practical Reason* (1788). It corresponds to the desire-to-act mode of consciousness. In modeling field theory, it is described mathematically by *adaptation laws*.

Recognition Engineering application areas and mathematical algorithms that classify patterns in data as belonging to particular classes or types.

Regression, linear A basic and widely used mathematical prediction method in the presence of uncertainty that combines probabilistic and deterministic aspects. Establishes a linear relationship among variables.

Schema Within CAS and genetic algorithm theory it is a mathematical notion corresponding to a generalized concept of a collection of building blocks of internal models (gene alleles, or substrings) that coevolves in the evolution process. Schemata are not used in genetic algorithms, but for their mathematical analysis. Schemata can be viewed as fuzzy submodels obtained by averaging nonfuzzy individual-agent models over a population. The evolution of schemata is more similar to MFT adaptation than learning at an individual agent level. CAS agents are nonfuzzy and nonadaptive. CAS schemata are fuzzy and adaptive. The genetic mechanism of preferential reproduction for better fitted agents creates a gradient in the space of parameters of fuzzy schemata leading to schemata adaptation.

Semantics In semiotics: relations between sign-vehicles and their designata. Its mathematical description in MFT is given by Judgment, or similarity measures relating input data and models.

Semiosis A process of *symbol* formation. It involves syntactics, semantics, and pragmatics. In Chapter 10, the MFT adaptation process is identified with the dynamic process of semiosis combining internal representations, meaning, and behavior.

Semiotics A science of signs, *symbols*, and their interpretations. Among founders are Peirce (1935–66) and Morris (1971).

Sensor fusion An engineering area of applications of intelligent algorithms. Combines data from multiple sensors into a unified picture of the world. Includes data *association* and sensor management, which is related to the attention function.

Shannon–Einsteinian similarity Measures the amount of information in the model about the world. Suitable for approximate models.

Sign A nonadaptive entity designating something else, such as a name designating a class of objects.

Signal and image processing Engineering application areas and mathematical algorithms that deal with signals and images. Includes spectrum estimation.

Sign-vehicle In semiotics: the media used as a sign. Its mathematical description in MFT is given by input data associated with internal models.

Similarity A measure of correspondence between the data and model. A similarity measure depends on selecting a proper data subset (segmentation, association) that is a combination of samples or pixels. This leads to *combinatorial explosion* in classical approaches. An *AZ-similarity* measure was developed in Chapter 4 that eliminates *combinatorial explosion* by using *fuzzy logic.* Also see *pixel similarity* and *partial similarity*. The similarity measure corresponds to the *Judgment* faculty in *Kant's theory.*

Spiritual substance Descartes separated spiritual phenomena from material ones and declared them to be properties of the two substances. Each one should be explained from its own principles and one should not be used to explain another. Descartes freed matter from materialized residues of the idea of emanation and created a condition for the development of science. Newton was disappointed that he could not surpass Descartes in this regard: his physics did not encompass spiritual substance. Today, we are close to creating a physics of spiritual substance: a mathematical theory of mind that corresponds to our physical intuition.

State parameters Model parameters in Kalman filter formalism.

Strong AI A term introduced by Searle: a belief that the material and energetic structure of the brain plays no principal role in the theory of intellect.

Symbol A dynamic entity, an adaptive emergent concept formed in the process of *semiosis*. In Chapter 10, the *MFT* agents (submodels) are identified with the dynamic symbols. An MFT symbol is a vortex of thinking, feeling, and acting. This definition is different from the often used notion of a static symbol, which is just a sign, as in "symbolic AI." Our definition corresponds to the semiotic analysis, analytical psychology of Jung, and Pribram's analysis of neural interactions in the brain. According to Jung, symbol is a dynamic entity relating consciousness and unconscious, a creative process.

Symbolic AI The often used designation for computational intelligence methods utilizing logical rule systems. This is a misnomer, as logical rule systems use *signs*, not *symbols*.

Syntactics In semiotics: relations among sign vehicles. Its mathematical description in MFT is given by a logic governing Understanding, or relationships among internal models.

Synthetic judgments a priori In *Kant's theory,* nontrivial (synthetic, nontautological) conclusions derived from fundamental (a priori) truths in such a way that their validity is undoubted (that is, of a priori origin). According to Kant, explaining the mind's ability to form synthetic judgments a priori is the main problem of the analysis of Pure Reason. He explained it as a special a priori ability. In *modeling field theory* this ability is due to hierarchical structures of the a priori model.

Target tracking In general: techniques of detecting objects and estimating their trajectories in sensory data. In this book: algorithms for the detection and prediction of multiple, possibly overlapping, moving patterns. Tracking algorithms are model based; they combine deterministic models with *uncertainty*, and they combine *apriority* of the motion models with *adaptivity* to trajectory parameters.

Thing-in-itself In *Kantian theory,* the infinite, forever impenetrable mysterious nature of an object.

Thinking A vortex of thought, a process of perception or cognition, involving concepts (internal models), emotions (similarity), and adaptation. Compare to *symbol*. Although used to denote a part of this process, the ability to operate with concepts.

Time and space According to Kant, time and space are prior to any experience, but he was ambivalent concerning the apriority of their origin. According to Jung, the primordial *archetypes* are timeless and spaceless (that is, time and space are of a more recent origin than the archetypes). The relatively recent origins of the concepts of time and space are indicated by observations of psychologists (Lèvy-Bruhl, 1910) and linguists (Whorf, 1936) concerning different conceptions of time and space in different peoples; in particular, some primitive tribes perceive time and space as not quite ordered globally with a higher level of the local orderliness (in the local region of time and space where the tribe currently lives). In the modeling field theory of mind, the concepts of time shall be explained due to the properties of a priori models.

Time travel Combined effects of quantum and relativistic theories could lead to the possibility of back-and-forth time travel. This is due to a nonzero probability of a space–time state that curves so much that a closed time-like line appears. If closed time-like lines could be exploited in a computer, there is a possibility that the computer has at its disposal the results of computations, before the computation began. According to Penrose (1989) this opens the possibility for noncomputable physics, which is needed to explain consciousness. The discovery of these new principles will constitute a theory that he calls Correct Quantum Gravitation.

Top-down processing A mathematical technique of recognizing predefined classes and concepts (top level) in the data (bottom level). Related to the principle of apriority and the philosophy of *realism*.

Turing's test A test that introduced a computational metaphor for the mind. It consisted in a thought experiment: a computer or a human is placed in a closed room and communications (questions and answers) are transmitted, say by a teletype. If as a result of such an interaction, one cannot tell if there is a human or a computer in the room, then the mind is similar to a computation.

Turing theory Formulated the concept of algorithmic computability and established its limitations related to Gödel's theory.

Uncertainty Mathematical methods of describing uncertainty include the theory of probability, the theory of chaos, and fuzzy logic. These theories are significantly interrelated. The theory of probability describes chances or relative frequencies of events that occur many times with random variations. Fuzzy logic describes events whose uncertainty is nonrandom and related to insufficient knowledge rather than random chance. The theory of chaos describes chaotic dynamics of certain processes. Uncertainties considered in this book are related to events and states, rather than dynamics.

Unconscious Psychic contents inaccessible to consciousness. The unconscious is known through scientific deductions. Unconscious contents can be classified into two general groups: personal and impersonal. The personal unconscious comprises life experiences that were forgotten or subliminally perceived, thought, or felt. Impersonal unconscious contents originate in the inherited possibilities of the psychic functioning in general. A significant part of these contents is common to all of humankind; Jung called these

the collective unconscious. Specific contents of the unconscious consist of archetypes or internal models.

Understanding A faculty of concepts, a source of general notions in Kantian theory. Described in the *Critique of Pure Reason* (1781). It corresponds to the thinking mode of consciousness. In modeling field theory, it is described mathematically by *internal models*. Penrose (1989) designates by "understanding" a specific awareness of the entire mathematical or physical theory. I call this intuition; some aspects can be understood the properties of internal representations or models (*understanding*); other aspects are related to the expansion of the internal models and represent a challenge to our contemporary rational understanding.

Universals General concepts of mind. In philosophical discussions, the issue of *adaptivity vs. apriority* is often referred to as "the origin of universals."

BIBLIOGRAPHY

Ackley, D.H., Hinton, G.E., and Sejnowski, T.J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science* **9**, 147–169.

Adelman, G. (1987). *Encyclopedia of Neuroscience.* Birkhaüser, Boston, MA.

Akaike, H. (1970). Statistical predictor identification. Ann. Stat. Math. **22**, 203–217.

Albus, J.S. (1991). *An outline for a theory of intelligence. IEEE Trans. Syst., Man Cyber.* **21**(3), 473–509.

Anderberg, M.R. (1973). *Cluster Analysis for Applications*. Academic Press, New York, NY.

Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed. John Wiley, New York.

Aristotle. (IV BC). *The Complete Works of Aristotle*, J. Barnes, ed. Bollingen Series, Princeton, NJ, 1995.

Aristotle. (IV BC). *Metaphysics*, W.D. Ross, trans. Bollingen Series, Princeton, NJ, 1995.

Aristotle. (IV BC). *De Anima (On the Soul)*, J.A. Smith, trans. Bollingen Series, Princeton, NJ, 1995.

Aristotle. (IV BC). *Topics*, W.A. Pickard-Cambridge, trans. Bolingen Series, Princeton, NJ, 1995.

Aquinas, T. (XIII) *Summa contra Gentiles*, English Dominican Fathers, trans. London, 1924.

Avicenna. (XI AD). *Kitab al-Shifa*. In *Avicenna*, S.M. Afnan, trans. George Allen & Unwin, London, 1958.

Baggenstoss, P.M. (1997). Joint density and structural learning by feature grouping. *Proceedings of the Conference on Intelligent Systems and Semiotics, ISAS'97*, pp. 124–129. Gaithersburg, MD.

Bar-Shalom, Y., and Li, X.-R. (1995). *Multitarget-Multisensor Tracking: Principles and Techniques*. Artech House, Norwood, MA.

Bar-Shalom, Y., and Tse, E. (1975). Tracking in a cluttered environment with probabilistic data association. *Automatica* **11**, 451–460.

Bartlett, M.S. (1946). On the theoretical specifications and sampling properties of autocorrelated time series. *J. Royal Statis. Soc. Supp.* **8**, 27–41.

Bayes, Rev. T. (1763). An essay toward solving a problem in the doctrine of chances. *Phil. Trans. R. Soc.* **53**, 370–418. Reprinted in *Biometrika* **45**, 293–315.

Bellman, R.E. (1961). *Adaptive Control Processes.* Princeton University Press, Princeton, NJ.

Berdyaev, N. (1969). The meaning of history (Smysl istorii, Russian). YMCA Press, Paris, France.

Berwick, R.C. (1982). *Locality principles and the acquisition of syntactic knowledge.* Ph.D. Thesis, MIT, Cambridge, MA.

Bhatnagar, R., and Kanal, L.N. (1993). Structural and probabilistic knowledge for abductive reasoning. *IEEE Trans. Pattern Anal. Machine Intell.* **15** (3), 233–245.

Bilbro, G.L., and Snyder, W.E. (1988). Range image restoration using mean field annealing. *IEEE Conference on Neural Information Processing Systems*, Denver, CO.

Bishop, C.M. (1996). *Neural Networks for Pattern Recognition*. Oxford University Press, NY.

Blackman, S.S. (1986). *Multiple Target Tracking with Radar Applications*. Artech House, Norwood, MA.

Blum, A.L., and Rivest, R.L. (1992). Training a 3-node neural networks is NP-complete. *Neural Networks* **5**, 117–127.

Bonnisone, P.P., Henrion, M., Kanal, L.N., and Lemmer, J.F. (1991). *Uncertainty in Artificial Intelligence 6*. North Holland, Amsterdam, The Netherlands.

Booker, H.G., Tao J-W., and Behroozi, Toosi, A.B. (1987). A scintillation theory of fading in long distance HF ionospheric communications. *J. Atmos. Terr. Phys.* **49**, 939–958.

Botha, R.P. (1991). *Challenging Chomsky. The Generative Garden Game.* Basil Blackwell, Oxford.

Bourgeois, F. and Lassalle, J.-C (1971) Algorithm for the assignment problem. *Commun. ACM* **14**, 805–806.

Broca, P.P. (1878). In *Encyclopedia of Neuroscience*, G. Adelman, ed. Birkhauser, Boston, MA, 1987, p. 589.

Brockett, R.W. (1991). Dynamical systems that sort lists, diagonalize matrices, and solve linear programming problems. *Linear Algebra Appl.* **146**, 79–91.

Brooks, R.A. (1983). Model-based three-dimensional interpretation of two-dimensional images. *IEEE Trans. Pattern Anal. Machine Intell.* **5**(2), 140–150.

Burdick, B.J., and Perlovsky, L.I. (1991). *Application of MLANS to Real-Time Learning of Targets.* Presidential Scientific Advisory Group Meeting, Washington, DC.

Burg, J.P. (1967). Maximum entropy spectral analysis. *Proceedings of the 37th Meeting of the Society for Exploration Geophysicists*, Dallas, TX.

Califano A., and Mohan, R. (1994). Multidimensional indexing for recognizing visual shapes. *IEEE Trans. Pattern Anal. Machine Intell.* **16**(4), 373–392.

Carpenter, G.A. (1989). Neural network models for pattern recognition and associative memory. *Neural Networks* **2**, 243–257.

Carpenter, G.A. (1994). A distributed outstar network for spatial pattern learning. *Neural Networks* **7**, 159–168.

Carpenter, G.A. (1996). Distributed ART networks for learning, recognition, and prediction, *Proceedings of the World Congress on Neural Networks (WCNN'96)*, pp. 333–344.

Carpenter, G.A. (1997). Distributed learning, recognition, and prediction by ART and ARTMAP neural networks. *Neural Networks* **10**, 1473–1494.

Carpenter, G.A., and Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Comput. Vision, Graphics Image Process.* **37**, 54–115.

Carpenter, G.A., Grossberg, S., and Reynolds, J.H. (1991a). ARTMAP. *Neural Networks* **4**, 565–588.

Carpenter, G.A., Grossberg, S., and Rosen, D.B. (1991b). Fuzzy ART. *Neural Networks* **4**, 759–771.

Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J.H., and Rosen, D.B. (1992). Fuzzy ARTMAP. *IEEE Trans. Neural Networks* **3**, 698–713.

Carpenter, G.A., Gjaja, M.N., Gopal, S., Markuzon, N., and Woodcock, C.E. (1997). ART and ARTMAP neural networks for applications: Self-organizing learning, recognition and prediction. In *Soft Computing Techniques in Knowledge-Based Intelligent Systems in Engineering*, L.C. Jain, ed., pp. 279–317. Springer-Verlag, New York.

Carpenter, G.A., Milenova, B.L., and Noeske, B.W. (1998). dARTMAP. *Neural Networks* **11** (5), pp. 793–813.

Chalmers, D. (1994). Facing up to the problem of consciousness. In *Explaining Consciousness: The Hard Problem*, ed. J. Shear. MIT Press, 1997.

Chapman, D. (1987). Planning for conjunctive goals. *Artificial Intelligence*. **32**, 333–377.

Chen, R.T., and Dyer, C.R. (1986). Model-based recognition in robotic vision. *ACM Comput. Surveys* **18**, 67–108.

Cherkassky, V., and Mulier, F. (1998). Learning from data: concepts, theory, and methods. J. Wiley & Sons, New York, NY.

Chomsky, N. (1972). *Language and Mind.* Harcourt, Brace Jovanovich, New York.

Chomsky, N. (1981). Principles and parameters in syntactic theory. In *Explanation in Linguistics. The Logical Problem of Language Acquisition*, N. Hornstein and D. Lightfoot, eds. Longman, London.

Cramer, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.

Crevier, D. (1993). *AI: The Tumultuos History of the Search for Artificial Intelligence*. Basic Books, New York, NY.

Dalai Lama, XIV. (1993). *The Buddhism of Tibet*. Ed. J. Hopkins. Snow Lion Pubs. Ithaca, NY.

Daugman, J.G. (1988). Brain metaphor and brain theory. In *Computational Neuroscience*, E. Schwartz, ed. MIT Press, Cambridge, MA.

Daum, F.E. (1990). Bounds on performance for multiple target tracking. *Proc. IEEE Trans. Automatic Control* **35**, pp. 833–839. (4).

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc.* **39B**, 1–38.

Descartes, R. (1637). Discourse on method. In *Discourse on Method and the Meditations*, F.E. Sutcliffe, trans. Penguin Books, New York, 1968.

Descartes, R. (1644). *Principles of Philosophy*, V.R. Miller and R.P. Miller, trans. D. Reidel, Dordrecht, The Netherlands, 1983.

Dixon, J. (1943). In Foreword to Kant (1781), *Critique of Pure Reason*. J. Wiley & Sons, New York, NY.

Dmitriev, V.A., and Perlovsky, L.I. (1996). Art form as an object of cognitive modeling (towards development of Vygotsky's semiotics model). *Proceedings of the Conference on Intelligent Systems and Semiotics '96*, Gaithersburg, MD, vol. 2, pp. 385–389.

Dmitriev, V.A., and Perlovsky, L.I. (1997). Cyberaesthetics and zoosemiology. *Proceedings of the Conference on Intelligent Systems and Semiotics '97*, Gaithersburg, MD, pp. 333–337.

Duda, R.O., and Fossum, H. (1966). Pattern classification by iteratively determined linear and piecewise linear disciminant functions. *IEEE Trans. Electr. Comp.* (**EC-15**), 220–232.

Duda, R.O., and Hart, P.E. (1973). *Pattern Classification and Scene Analysis*. John Wiley, New York.

Durbin, J. (1960). The fitting of time-series models. *Internation Stati. Rev.* **28**, 233–244.

Einstein, A., and Hopf. L. (1910). Uber einen Satz der Wahrscheinlichkeitsrechnung und seine Anwendung an die Strahlungstheorie. *Ann. Phys.* **33**, 1096–1104.

Feynman, R.P. (1972). *Statistical Mechanics.* Benjamin, Reading, MA.

Feynman, R.P. (1982). Simulating physics with computers. *Foundations of Physics*, **16**, pp. 507–531.

Feynman, R.P. (1986). Quantum mechanical computers. *Int. J. Theor. Physics* **21**(6/7), 467–488.

Fogel, D.B. (1995). Evolutionary Computation: Toward a new philosophy of machine intelligence. *IEEE Press*, Piscataway, NJ.

Fortmann, T.E., Bar-Shalom, Y., Scheffe, M., and Gelfand, S. (1985). Detection thresholds for tracking in clutter—A connection between estimation and signal processing. *IEEE Trans. Automatic Control* **AC-30**(3), 221–229.

Franchi, P.R., and Tichovolsky, E.J. (1989). *Phase Screen Modulation as a Source of Clutter Related Noise in Over-the-Horizon Radars*, RADC-TR-89-296.

Franklin, S.P. (1995). *Artificial Minds*. MIT Press.

Frank-Kamenetsky, M. (1997). *Unraveling DNA: The Most Important Molecule of Life*. Addison-Wesley, Reading, MA.

Freeman, W.J. (1975). *Mass Action in the Nervous System.* Academic Press, New York.

Freeman, W.J. (1996). A biological model for construction of meaning to serve as an interface between intelligent system and its environment. *ISAS'96, Intelligent Systems: A Semiotic Perspective*, Gaithersburg, MD.

Freud, S. (1895). *Project for a Scientific Psychology*. In *The Standard Edition of the Complete Psychological Works*, J. Strachey, trans., Vol. 1, pp. 283–398, Hogart Press, London.

Freud, S. (1900). *Interpretation of Dreams.* In *The Standard Edition of the Complete Psychological Works*, J. Strachey, trans., Vol. 2, Hogart Press, London.

Fukunaga, K. (1972). *Introduction to Statistical Pattern Recognition.* Academic Press. New York; 2nd ed. 1991.

Fukunaga K., and Hayes, R.R. (1988). Statistical classifier design and evaluation, *Purdue University Report TR-EE 88-19*, West Lafayette, IN.

Fukushima K., and Miyake, S. (1982). Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Pattern Recognition* (**15**), p. 445–451.

Gardner, H. (1987). *The Mind's New Science*. Harper Collins, New York.

Garvin, L., and Perlovsky, L.I. (1995). Statistical quantum theory and statistical pattern recognition. In *Current Topics in Pattern Recognition Research*, S.G. Pandalai, ed. Research Trends, Trivandrum, India.

Garvin, L., Marzetta, T., and Perlovsky, L. (1989). Fusion and effects of fusion on exoatmospheric discrimination. *Proceedings of 1989 Tri-Service Data Fusion Symposium*, San Diego, CA.

Girosi, F., Jones, M., and Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural Comp.* **7**(2), 219–269.

Goj, W.W. (1993). *Synthetic Aperture Radar and Electronic Warfare.* Artech House, Norwood, MA.

Goldfarb, L. (1996). Inductive learning as a central learning process. *ISAS'96, Intelligent Systems: A Semiotic Perspective*, Gaithersburg, MD.

Gödel, K. (1986). *Kurt Gödel Collected Works*, I, S. Feferman at al., eds. Oxford University Press, New York.

Goodman, L.E. (1977). *Rambam, Readings in the Philosophy of Moses Maimonides.* Schocken Books, New York.

Graham, M.L., and Streit R.L. (1994). The Cramer-Rao bound for multiple target tracking algorithms. *NUWC-NPT Tech. Rep.* 10,406, 969–970; **14**(2), 97–98.

Greineder, S. (1995). Private communication.

Grimson, W.E.L., and Huttenlocher, D.P. (1991). Introduction to the special issue on interpretation of 3-D scenes. *IEEE Trans. Pattern Analy. Machine Intelligence* **13**(10), 969–970; **14**(2), 97–98.

Grimson, W.E.L., and Lozano-Perez, T. (1984). Model-based recognition and localization from sparse range or tactile data. *Int. J. Robotics Res.* **3**(3), 3–35.

Grossberg, S. (1968). Some nonlinear networks capable of learning a spatial pattern of arbitrary complexity. *Proc. Nat. Acad. Sci. U.S.A.* **59**, 368–372.

Grossberg, S. (1970). Neural pattern discrimination. *J. Theoret. Biol.* **27**, 291–337.

Grossberg, S. (1971). Embedding fields: Underlying philosophy, mathematics and applications to psychology, physiology, and anatomy. *J. Cybern.* **1**, 28–50.

Grossberg, S. (1972a). A neural theory of punishment and avoidance, I: Qualitative theory. *Math. Biosci.* **15**, 39–67.

Grossberg, S. (1972b). A neural theory of punishment and avoidance, II: Quantitative theory. *Math. Biosci.* **15**, 253–285.

Grossberg, S. (1975). A neural model of attention, reinforcement and discrimination learning. In *International Review of Neurobiology*, C.C. Pfeiffer, ed. Vol. 18, Academic Press, New York.

Grossberg, S. (1976). Adaptive pattern classification and universal recording: II. Feedback, oscillation, olfaction, and illusions. *Biol. Cybern.* **23**, 187–207.

Grossberg, S. (1978). A theory of human memory: Self-organization and performance of sensory-motor codes, maps and plans. In *Progress in Theoretical Biology*, R. Rosen and F. Snell, eds., Vol. 5, pp. 233–374. Academic Press, New York.

Grossberg, S. (1980a). How does a brain build a cognitive code? *Psychol. Rev.* **87**, 1–51.

Grossberg, S. (1980b). Biological competition: Decision rules, pattern formation, and oscillations. *Proc. Natl. Acad. Sci. U.S.A.* **77**, 2338–2342.

Grossberg, S. (1982). *Studies of Mind and Brain.* D. Reidel, Dordrecht, Holland.

Grossberg, S. (1983). The quantized geometry of visual space: The coherent computation of depth, form, and lightness, *Behav. Brain Sci.* **6**, 625–692.

Grossberg, S. (1988a). *Neural Networks and Natural Intelligence*. MIT Press, Cambridge, MA.

Grossberg, S. (1988b). Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks* **1**(1), 17–61.

Grossberg, S. (1995). The attentive brain. *Am. Sci.* **83**, 483–449s.

Grossberg, S., and Gutowski, W.E. (1987). Neural dynamics of decision making under risk: Affective balance and cognitive-emotional interactions. *Psychol. Rev.*, **94**(3), 300–318.

Grossberg, S., and Kuperstein, M. (1989). *Neural Dynamics of Adaptive Sensory-Motor Control*. Pergamon Press, New York.

Grossberg, S., and Levine, D.S. (1987). Neural dynamics of attentionally modulated Pavlovian conditioning: Blocking, inter-stimulus interval, and secondary reinforcement. *Psychobiol.* **15**(3), 195–240.

Grossberg, S., and Merrill, J.W.L. (1992). A neural network model of adaptively timed reinforcement learning and hippocampal dynamics. *Cog. Brain Res.* **1**, 3–38.

Grossberg, S., and Schmajuk, N.A. (1987). Neural dynamics of attentionally modulated Pavlovian conditioning: Conditioned reinforcement, inhibition, and opponent processing. *Psychobiology* **15**(3), 195–240.

Grossberg, S., and Schmajuk, N.A. (1989). Neural dynamics of adaptive timing and temporal discrimination during associative learning. *Neural Networks* **2**, 79–102.

Hall, D.L. (1997). An introduction to multisensor data fusion. *IEEE Proc.* **85**(1), .6–23.

Hall, D.L., and Linn, R.J. (1990). A taxonomy of algorithms for multisensor data fusion. *Proc. Tri-Service Data Fusion Symp.* 13–29.

Hall, D.L., Linn, R.J. and Llinas, J. (1991). A survey of data fusion systems. *Proc. SPIE Conf. Data Structure and Target Classification*. SPIE **1470**, 13–36. Orlando, FL.

Hameroff, S.R. (1987). *Ultimate Computing. Biomolecular Consciousness and Nanotechnology.* North-Holland, Amsterdam.

Hameroff, S.R. (1994). *Toward a Scientific Basis for Consciousness*. MIT Press, Cambridge, MA.

Haykin, S. (1998). *Neural Networks*, 2nd ed. Prentice-Hall, Englewood Cliffs, NJ.

Hebb, D. (1949). *Organization of Behavior*. John Wiley, New York.

Hegel, G.W.F. (1979). *Phenomenology of Spirit*. Tr. A.V. Miller. Oxford University Press, New York, NY.

Hildebrand, F.B. (1952). *Methods of Applied Mathematics.* Prentice-Hall, Englewood Cliffs, NJ.

Ho, Y.C., and Agrawala, A. (1968). On pattern classification algorithms: Introduction and survey. *IEEE Proc.* **56**, 2101–2114.

Holland, J.H. (1992). *Adaptation in Natural and Artificial Systems.* MIT Press, Cambridge, MA.

Holland, J.H. (1995). *Hidden Order.* Addison-Wesley, Reading, MA.

Hopfield, J. (1982). Neural networks and physical systems with emerging collective computational abilities. *Proc. Natl. Acad. Sci.* **79**, 2554–2558.

Horgan, T., and Tienson, J. (1989). Representations without rules. *Philosophical Topics*. **17**, 147–174.

Jacobs, R.A. Jordan, M.I., Nowlan, S.J., and Hinton, G.E. (1991). Adaptive mixtures of local experts. *Neural Comp.* **3**, 79–87.

Jaffer, A.J., and Bar-Shalom, Y. (1972). On optimal tracking in multiple target environments. *Proceedings of the 3rd Symposium Non-Linear Estimation Theory and Its Applications*,San Diego, CA, pp. 112–117.

James, W. (1890). *The Principles of Psychology*. Dover Books, New York, 1950.

Jaynes, E.T. (1957). Information theory and statistical mechanics I, *Phys. Rev.* **106**, 620–630.

Jaynes, E.T. (1994). *Probability Theory: The Logic of Science*. John Wiley, New York.

Jaynes, J. (1976). *The Origin of Consciousness in the Breakdown of the Bicameral Mind*. Houghton Mifflin, Boston, MA.

Jibu, M., and Yasue, K. (1993). The basics of quantum brain dynamics. In *Rethinking Neural Netwroks*, K.H. Pribram, ed., pp. 121–145. Lawrence Erlbaum, Hillsdale, NJ.

Jordan, M.I., and Jacobs, R.A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Comp.* **6**, 181–214.

Jung, C.G. (1921). *Psychological Types.* In *The Collected Works*, Vol. 6, Bollingen Series XX, Princeton University Press, Princeton, NJ, 1971.

Jung, C.G. (1934). *Archetypes of the Collective Unconscious.* In *The Collected Works*, Vol. 9, part II. Bollingen Series XX, Princeton University Press, Princeton, NJ, 1969.

Jung, C.G. (1951). *Aion, Researches into the Phenomenology of the Self*, 2nd ed. In *The Collected Works*, Vol. 9, part II, Bollingen Series XX. Princeton University Press, Princeton, NJ, 1969.

Kalman, R.E. (1960). A new approach to linear filtering and prediction. *Journal of Basic Engineering*, pp. 35–46.

Kant, I. (1781). *Critique of Pure Reason*, J.M.D. Meiklejohn, trans. John Wiley, New York, 1943.

Kant, I. (1787). Letter to Reingold. In *Tractats and Letters*, Moscow, 1980.

Kant, I. (1788). *Critique of Practical Reason*, J.H Bernard, trans. Hafner, New York, 1986.

Kant, I. (1790). *Critique of Judgment*, J.H. Bernard, trans., 2nd ed. Macmillan, London, 1914.

Kay, S.M., and Marple, S.L., Jr. (1981). Spectrum analysis—A modern perspective. *IEEE Proc.* **69**(11), 1380–1419.

Keshavan, H.R., Barnett, J., Geiger, D., and Verma, T. (1993). Introduction to the special section on probabilisitc reasoning. *IEEE Trans. PAMI*, **15**(3), 193–195.

Klein, L.A. (1993). Sensor and data fusion concepts and applications. *SPIE Press*, **14**.

Klopf, A.H. (1987). *A Neuronal Model of Classical Conditioning*. Air Force Wright Aeronautical Laboratories Report AFWAL-TR-87-1139, WPAFB, OH.

Kohonen, T. (1984). *Self-Organization and Associative Memory*. Springer-Verlag, New York.

Koster, J., and May, R. (1981). *Levels of Syntactic Representation*. Foris Publications, Dordrecht, The Netherlands.

Kryukov, V.I. (1988). Short-term memory as a metastable state: "Neurolocator," a model of attention, *IEEE Conference on Neural Information Processing Systems*, Denver, CO.

Kullback, S., and Leibler, R.A. (1951). On information and sufficiency. *Ann. Math. Stat.* **22**, 76–86.

Kumar, V.P., and Manolakos, E.S. (1996). Combining soft decision algorithms and scale-sequential hypotheses pruning for object recognition. *Proceedings of the Conference on Intelligent Systems and Semiotics '96*, Gaithersburg, MD, Vol. 1, pp. 215–220.

Lakatos, I., and Musgrave, A.E. (1970). Criticism and the Growth of Knowledge. *Cambridge University Press*, Cambridge.

Lakoff, G., and Johnson, M. (1983). *Metaphors We Live By*. University of Chicago Press, Chicago, IL.

Lamdan, Y., and Wolfson, H.J. (1988). Geometric hashing: A general and efficient recognition scheme. *Proceedings of the 2nd International Conference Computer Vision.* Cambridge, MA. pp. 143–148.

Landau, L.D., and Lifshitz, E.M. (1980). *Statistical Physics*. Pergamon Press, London.

Lang, J.K., Waibel, A.H., and Hinton, G.E. (1990). A time-delay neural network architecture for isolated word recognition. *Neural Networks*. **3**(1), 23–43.

Levine, D.S. (1994). Steps toward a neural network theory of self-actualization. *World Congress on Neural Networks*, San Diego, Vol. 1, pp. 215–220. Erlbaum, Mahwah, NJ.

Levine, D.S. (1995). Do we know what we want? *World Congress on Neural Networks*, Washington, Vol. II, pp. 955–962. Erlbaum, Mahwah, NJ.

Levine, D.S. (1996). Neural modeling of prefrontal executive function. *Proceedings of the Conference on Intelligent Systems and Semiotics '96*, Gaithersburg, MD, Vol. 1, pp. 240–46.

Levine, D.S. (in preparation). *Common Sense and Common Nonsense*. New York: Oxford University Press, New York.

Levine, D.S., Leven, S.J., and Prueitt, P.S. (1992). Integration, disintegration, and the frontal lobes. In *Motivation, Emotion, and Goal Direction in Neural Networks*, D.S. Levine, and S.J. Leven, eds. Erlbaum, Hillsdale, NJ.

Levinson, N. (1947). the Wiener RMS error criterion in filter design and prediction. *J. Math. Phys.* **25**, 261–278.

Levitin, L.B., Schapiro, B., Perlovsky, L.I. (1996). Zipf's law revisited: Evolutionary model of emergent multiresolution classification. *Proceedings of the Conference on Intelligent Systems and Semiotics '96*, Gaithersburg, MD, Vol. 1, pp. 65–70.

Lèvy-Bruhl, L. (1910). Les Fonctions mentales dans les sociètès infèrieures, tr. L.A. Clare. In *How Natives Think*. London, 1926.

Li, X.R., and Bar-Shalom, Y. (1993). Design of an interactive multiple model algorithm for air traffic control tracking. *IEEE Trans. Control Syst. Technol.* **1**(3), 186–194.

Llinas, J., and Waltz, E. (1990). Multisensor data fusion. Artech House.

Longstaff, I.D., and Cross, J.F. (1987). A pattern recognition approach to understanding the multi-layer perceptron. *Pattern Recognition Lett.* **5**(5), 315–320.

Lorenz, K. (1981). *The Foundations of Ethology.* Springer Verlag, New York.

MacLean, P.D. (1952). In *Encyclopedia of Neuroscience*, G. Adelman, ed., p. 589. Birkhaüser, Boston, MA.

Maes, P. (1991). Designing Autonomous Agents. MIT Press, Cabridge, MA.

Maimonides, M. (1190). *The Guide for the Perplexed*, M. Friedlander, trans. Dover, New York, NY.

Malsburg, C. von der (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*. **14**, 85–100.

Marr D. (1982). *Vision*. Freeman, San Francisco, CA.

Marty, K. (1976) *Linear and Combinatorial Programming*. John Wiley, New York.

Marzetta, T.L. (1994). *A New Approach to Covariance Matrix Estimation When the Sample Covariance Matrix Is Singular*. Unpublished.

Marzetta, T.L. (1995). EM algorithm for estimating the parameters of a multivariate complex Rician density for polarimetric SAR. *Proceedings of the International Conference Acoustic, Speech, and Signal Processing*, Detroit, MI.

Mazo, R. (1994). Lament made visible: a study of paramusical elements in Russian lament. In *Themes and Variations*, ed. B. Yung and J.C. Lam. Harvard University Press.

McCamey, K.J., Pauli, M., Perlovsky, L.I., Marzetta, T.L., Martinsen, E.A., and Garvin, L.C. (1994). Detection of targets in a high clutter background. *IRIS Conference*, Monterey, CA.

McCulloch, W.S. (1961). *What Is a Number That a Man May Know It, and a Man, That He May Know a Number?* The 9th Alfred Korzybski Memorial Lecture. *Gen. Semant. Bull.* **26**, 17, 18. Also in McCulloch (1965).

McCulloch, W.S. (1965). *Embodiments of Mind.* 2nd ed. MIT Press, Cambridge, MA, 1988.

McCulloch, W., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **7**, 115–133.

Mehrotra, K.G., Mohan, C.K., and Ranka, S. (1991). Bounds on the number of samples needed for neural learning. *IEEE Trans. Neural Networks* **2**(6), 548–554.

Metropolis, N. (1980). New Directions in Physics. UCP, Los Alamos, NM.

Meystel, A. (1988). Intelligent control in robotics. *J. Robotic Syst.* **5**(4), pp. 269–308.

Meystel, A. (1995). *Semiotic modeling and Situational Analysis*. Ad. Rem, Bala Cynwyd, PA.

Meystel, A. (1996). *Intelligent Systems*. IFAC World Congress, San Francisco, CA.

Michalski, R.S., Carbonell, J.G., and Mitchell, T.M. (1986). *Machine Learning: An Artificial Intelligence Approach*, Vol. II. Morgan Kaufmann, Los Altos, CA.

Minsky, M.L. (1954). Neural-analog networks and the brain model problem, Ph.D. Thesis, Princeton University, Princeton, NJ.

Minsky, M.L. (1968a). *Semantic Information Processing.* MIT Press, Cambridge, MA.

Minsky, M.L. (1968b). Matter, mind, and models. In *Semantic Information Processing*, M.L. Minsky, ed. MIT Press, Cambridge, MA.

Minsky, M.L. (1975). A framework for representing knowlege. In *The Psychology of Computer Vision*, P. H. Whinston, ed. McGraw-Hill, New York.

Minsky, M. (1985). *Society of Mind*. MIT Press.

Minsky, M.L., and Papert, S.A. (1969, 1988). *Perceptrons*. MIT Press, Cambridge, MA.

Moore, B., and Poggio, T. (1988). Representation properties of multilayer feedforward networks. The 1st Annual INNS Meeting, Boston, MA., Pergamon Press.

Moravscik, E.A., and Wirth, J.R. (1980). *Syntax and Semantics, Vol. 13: Current Approaches to Syntax*. Academic Press, New York.

More, L.T. (1934). *Isaac Newton a Biography*. Cambridge University Press, Cambridge.

Morris C. (1971). *Writings on the General Theory of Signs*, Th. A. Sebeok, ed. Mouton, The Hague.

Munkres, J. (1957). Algorithms for the assignment and transportation problems. *J. SIAM* **5**, 32–38.

Muratet, M.A., Brahm, S., Schoendorf, W.H., Perlovsky, L.I., Burdick, B.J., Lash, M.E. (1998). Interceptor designation and discrimination with concurrent algorithms. 7th Ann. AIAA/BMDO Tech. Readiness Conf., San Diego, CA.

Nahi, N.E. (1969) Optimal recursive estimation with uncertain observations. *IEEE Trans. Inform. Theory* **IT-15**, 457–462.

Nagel, T. (1974). What is it like to be a bat? The Philosophical Review. **11**, 207–212.

Negahdaripour, S., and Jain, A.K. (1991). *Final Report of the NSF Workshop on the Challenges in Computer Vision Research; Future Directions of Research.* National Science Foundation, Washington, DC.

Neumann von, J. (1927). Density operator in quantum mechanics. *Nachr. Acad. Wiss.* **3**, 245–273.

Nevatia R., and Binford, T.O. (1977). Description and recognition of curved objects. *Artificial Intelligence* **8**(1), 77–98.

Newell, A. (1983). Intellectual issues in the history of artificial intelligence. In *The Study of Information*, F. Machlup and U. Mansfield, eds. John Wiley, New York.

Newell, A., and Simon, H.A. (1982). *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ.

Nietzsche, F. (1886). Beyond good and evil. In *Basic Writings of Nietzsche*, Kaufman, trans. Random House, New York, 1968.

Nilsson, N.J. (1965). *Learning Machines.* McGraw-Hill, New York.

Occam, W. (XIV). *Summa logicae*. In *Occam's Theory of Terms*, M.J. Loux, trans., 1974, and *Occam's Theory of Propositions*, A.J. Freddoso and H. Schuurman, trans. 1980. University of Notre Dame Press, Notre Dame, IN.

Oja, E. (1992). Principal components, minor components, and linear neural networks. *Neural Networks*. **5**, 927–935.

Olive, J.P., Greenwood, A., and Coleman J. (1993). *Acoustics of American English Speech.* Springer-Verlag, New York.

Papez, J.W. (1937). Emotions. In *Encyclopedia of Neuroscience*, G. Adelman, ed., p. 589. Birkhaüser, Boston, MA, 1987.

Parker, D.B. (1985). Learning logic. Tech. Rep. tr-47, Center Comp. Res. Econom. manag. Sc., MIT.

Parra-Loera, R., Thompson, W.E., and Akbar, S.A. (1992). Multilevel distributed fusion of multisesnsor data. *Conference on Signal Processing, Sensor Fusion and Target Recognition*, *SPIE Proceedings*, Vol. 1699.

Peirce, C.S. (1935–66). *Collected Papers of Charles Sanders Peirce*. Harvard University Press, Cambridge, MA.

Penrose, R. (1989). *The Emperor's New Mind*. Oxford University Press, Oxford, England.

Penrose, R. (1994). *Shadows of the Mind*. Oxford University Press, Oxford, England.
Perlovsky, L.I. (1980). On estimates of the parameters of the autoregressive process: Spectrum vs. autocorrelation function. *American Statistical Association Annual Meeting*, Houston, TX.
Perlovsky, L.I. (1987a). *Multiple Sensor Fusion and Neural Networks.* DARPA Neural Network Study, MIT/Lincoln Laboratory, Lexington, MA.
Perlovsky, L.I. (1987b). Robust estimates for parameters of oscillating systems. *SIAM 35th Anniversary Meeting*, Denver, CO.
Perlovsky, L.I. (1988a). Neural networks for sensor fusion and adaptive classification. *Proceedings of the 1st Annual International Neural Network Society Meeting*, Boston, MA, p. 42.
Perlovsky, L.I. (1988b). Cramer–Rao bounds for the estimation of means in a clustering problem. *Pattern Recog. Lett.* **8**, 1–3.
Perlovsky, L.I. (1988c). Frequency estimates for simple oscillating systems under random forcing. In *Linear Algebra in Signals, Systems & Control*, B.N. Datta et al., eds. SIAM Press, PA.
Perlovsky, L.I. (1989a). Cramer–Rao bounds for the estimation of normal mixtures. *Pattern Recog. Lett.*, **10**, 141–148.
Perlovsky, L.I. (1989b). Neural network for adaptive processing of quantum limited images. *SPIE Meeting on Quantum Limited Image and Information Processing II*, North Falmouth, MA.
Perlovsky, L.I. (1990a). Application of neural networks to transient signal classification. *Proceedings of the Government Neural Network Applications Workshop*, NOSC, San Diego, CA, p. 57.
Perlovsky, L.I. (1990b). ATR performance prediction model development using maximum likelihood artificial neural system. *ATR Systems & Technical Conference*, NWSC, Silver Springs, MD.
Perlovsky, L.I. (1991a). Tracking multiple objects in visual imagery. *Conference on Neural Networks for Vision and Image Processing*, Tyngsboro, MA.
Perlovsky, L. I. (1991b). Decision directed sensor fusion. *Proceedings of the 2nd Government Neural Network Applications Workshop*, Huntsville, AL, Sect. 1.6, pp. 1–5.
Perlovsky, L.I. (1992a). Track before detect using model based neural network. *Proceedings of the 3rd Government Neural Network Applications Workshop*, Dayton, OH, Vol. 1, pp. 85–88.
Perlovsky, L.I. (1992b). Model based sensor fusion. *Proceedings of SPIE Conference on Sensor Fusion*, Boston, MA, **1828**, pp. 197–200.
Perlovsky, L.I. (1993). MLANS neural network for sensor fusion. *Proceedings of the IEEE NAECON Conference*, Dayton, OH, pp. 880–884.
Perlovsky, L.I. (1994a). Computational concepts in classification: Neural networks, statistical pattern recognition, and model based vision. *J. Math. Imaging Vision*, **4** (1), 81–110.
Perlovsky, L.I. (1994b). A model based neural network for transient signal processing. *Neural Networks* **7**(3), 565–572.
Perlovsky, L.I. (1995). MLANS tracker. *Proceedings of the ARPA Counterdrug Over The Horizon Radar Workshop*, Chesapeake, VA, pp. A147–A173.
Perlovsky, L.I. (1996a). Model-based neural network for pattern recognition. *Proceedings of the World Congress on Neural Networks*, San Diego, CA, pp. 380–383. Erlbaum, Hillsdale, NJ.
Perlovsky, L.I. (1996b). Einsteinian neural network. *Proceedings of the World Congress on Neural Networks*, San Diego, CA, pp. 603–606. Erlbaum, Hillsdale, NJ.
Perlovsky, L.I. (1996c). Mathematical concepts of intellect. *Proceedings of the World Congress on Neural Networks*, San Diego, CA, pp. 1013–1016. Erlbaum, Hillsdale, NJ.
Perlovsky, L.I. (1996d). Aristotle, complexity and fuzzy logic. *Proceedings of the World Conference on Neural Networks*, San Diego, CA, p. 1152. Erlbaum, Hillsdale, NJ.
Perlovsky, L.I. (1996e). Statistical and predictive models in neural networks. *Proceedings of the World Conference on Neural Networks*, San Diego, CA, pp. 668–671. Erlbaum, Hillsdale, NJ.
Perlovsky, L.I. (1996f). Complexity of recognition: Aristotle, Gödel, Zadeh. *IFAC Triennial World Congress*, San Fracisco, CA.

Perlovsky, L.I. (1996g). Fuzzy logic of Aristotelian Forms. *Proceedings of the Conference on Intelligent Systems and Semiotics '96*, Gaithersburg, MD, Vol. 1, pp. 43–48.

Perlovsky, L.I. (1996h). Intelligence of recognition. Introduction. *Intelligent Systems and Semiotics '96*, Gaithersburg, MD.

Perlovsky, L.I. (1996i). Gödel theorem and semiotics. *Proceedings of the Conference on Intelligent Systems and Semiotics '96*, Gaithersburg, MD, Vol. 2, pp. 14–18.

Perlovsky, L.I. (1997a). Cramer–Rao bound for tracking in clutter and tracking multiple objects. *Pattern Recog. Lett.* **18**(3), 283–288.

Perlovsky, L.I. (1997b). Mathematical aspects of cyberaesthetics. *Proceedings of the Conference on Intelligent Systems and Semiotics '97*, Gaithersburg, MD, pp. 319–324.

Perlovsky, L.I. (1997c). Towards quantum field theory of symbol. *Proceedings of the Conference on Intelligent Systems and Semiotics '97*, Gaithersburg, MD, pp. 295–300.

Perlovsky, L.I. (1997d). Information, likelihood and Einstein. *Proceedings of the Conference on Intelligent Systems and Semiotics '97*, Gaithersburg, MD, pp. 307–312.

Perlovsky, L.I. (1997e). Modeling fields, evolutionary computations, and hierarchies. Plenary Talk. *Conference on Intelligent Systems and Semiotics '97*, Gaithersburg, MD.

Perlovsky, L.I. (1997f). Physical concepts of intellect. *Proc. Russian Acad. Sci.* **354**(3), 320–323.

Perlovsky, L.I. (1998a). *Conundrum of Combinatorial Complexity*. IEEE Trans. PAMI, **20**(6) 666–670.

Perlovsky, L.I. (1998b). *Cyberaesthetics: aesthetics, symbol, and control*. Proc. STIS'98: Joint Conf. Science & Tech. Intelligent Sys. (ISIC/CIRA/ISAS), Gaithersburg, MD.

Perlovsky, L.I. (1998c). *Models, Similarity, and Complexity*. Proc. STIS'98: Joint Conf. Science & Tech. Intelligent Sys. (ISIC/CIRA/ISAS), Gaithersburg, MD.

Perlovsky, L.I. (1998d). *Semiotics, Mind, and Architecture of Intelligent Target Tracker*. Proc. STIS '98: Joint Conf. Science & Tech. Intelligent Sys. (ISIC/CIRA/ISAS), Gaithersburg, MD.

Perlovsky, L.I. (1998e). *Computational Complexity and the Origin of Universals*. Proc. World Congress on Philosophy, Boston, MA.

Perlovsky, L.I., and Jaskolski, J.V. (1994e). Maximum likelihood adaptive neural controller. *Neural Networks* **7**(4), 671–680.

Perlovsky, L.I., and Marzetta, T.L. (1992). Estimating a covariance matrix from incomplete independent realizations of a random vector. *IEEE Trans. SP* **40**(8), 2097–2100.

Perlovsky, L.I., and McManus, M.M. (1991). Maximum likelihood neural networks for sensor fusion and adaptive classification. *Neural Networks* **4**(1), 89–102.

Perlovsky, L.I., and Plum, C.P. (1991). Model based multiple target tracking using MLANS. *Proceedings of the Third Biennial ASSP Mini Conference*, Boston, MA, pp. 56–59.

Perlovsky, L.I., and Van Gelder, P. (1980). The biological clock hypothesis and eye movements. *Society for Mathematical Psychology Annual Meeting*, Madison, WI.

Perlovsky, L.I., Schoendorf, W.H., Clark, L.G., and Keller, T.L. (1992). Information theoretic performance model for ATR. *IRIS Meeting on Passive Sensors*, Laurel, MD.

Perlovsky, L.I., Coons, R. P., Streit, R.L., Luginbuhl, T.E., and Greineder, S. (1994). Application of MLANS to signal classification. *J. Underwater Acoust.* **44** (2), 783–809.

Perlovsky, L.I., Chernick, J.A., and Schoendorf, W.H. (1995). Multi-sensor ATR and identification friend or foe using MLANS. *Neural Networks* **8** (7/8), 1185–1200.

Perlovsky, L.I., Schoendorf, W.H., Tye, D.M., and Chang, W. (1995b). Concurrent classification and tracking using maximum likelihood adaptive neural system. *J. Underwater Acoust.* **45**(2), 399–414.

Perlovsky, L.I., Plum, C.P., Franchi, P.R., Tichovolsky, E.J., Choi, D.S., and Weijers, B. (1997a). Einsteinian neural network for spectrum estimation. *Neural Networks*, **10**(9), pp. 1541–1546.

Perlovsky, L.I., Schoendorf, W.H., Garvin, L.C., Chang, W., and Monti J. (1997b). Development of concurrent classification and tracking. *J. Underwater Acoust.* **47**(1), 202–210.

Perlovsky, L.I., Schoendorf, W.H., Burdick, B.J., and Tye, D.M. (1997c). Model-based neural network for target detection in SAR images. *IEEE Trans. Image Proc.* **6**(1), 203–216.

Perlovsky, L.I., Webb, V.H., Bradley, S.R., and Hansen, C.A. (1998a). Improved ROTHR detection and tracking using MLANS. *AGU Radio Sci.* **33**(4), pp. 127–147.

Perlovsky, L.I., Webb, V.H., Plum, C.P., Franchi, P.R., Tichovolsky, E.J., and Weijers, B. (1998b). Analysis of spread-doppler clutter observed in OTH radar spectra. A.F. Geoph. Lab. Report T-89-31, Lincoln, MA.

Pitts, W., and McCulloch, W.S. (1947). How we know universals: The perception of auditory and visual forms. *Bull. Math. Biophys.* **9**, 127–147.

Plato. (IV BC). *Parmenides*. In *Plato*, L. Cooper, trans. Oxford University Press, New York.

Plato. (IV BC). *Complete Works*. J. Cooper, ed. Hackett Publishing, Indianapolis, IN, 1997.

Plotinus. (250). *The Six Enneads*. J. Dillon, ed. S. MacKenna, trans. Penguin, New York, NY, 1991.

Poggio, T. (1988). Learning, Regularization and Splines. The 1st Annual INNS Meeting, Vol. 1, p. 211.

Postaire, J.B., and Vasseur, C.P.A. (1981). An approximate solution to normal mixture identification with applications to unsupervised pattern classification, *IEEE PAMI* **3**(2), 163–179.

Press, S.J. (1989). *Bayesian Statistics: Principles, Models, and Applications*. John Wiley, New York.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. (1989). *Numerical Recipes*. Cambridge University Press, New York.

Pribram, K. (1971). *Languages of the Brain*. Prentice-Hall, Englewood Cliffs, NJ.

Pribram, K. (1993). *Rethinking Neural Networks: Quantum Fields and Biological Data*. Erlbaum, Hillsdale, NJ.

Priestly, M.B., (1981). *Special Analysis and Time Series*. Academic Press, New York, NY.

Proakis, J.G., Rader, C.M., Ling, F., and Nikias, C.L. (1992). *Advanced Digital Signal processing*. Macmillan, New York, NY.

Putnam, H. (1995). Review of *Shadows of the Mind*, by R. Penrose. Bull. Am. Math. Soc., **32**(5), 370–373.

Reilly, D.L., Scofield, C.L., Elbaum, C., and Cooper, L.N. (1987). Learning System Architectures Composed of Multiple Learning Modules. 1st Int. Conf. Neural Networks, IEEE.

Riasanovsky, N.V. (1992). *The Emergence of Romanticism*. Oxford University Press, New York, NY.

Ricciardi, L.M., and Umezawa, H. (1967). *Kybernetik* **4**, 44.

Rice, S.O. (1944). Mathematical analysis of random noise. *Bell Syst. Tech. J.* **23**, 282–332; reprinted in Wax (1954).

Rice, S.O. (1945). Mathematical analysis of random noise II. *Bell Syst. Tech. J.* **24**, 46–156; reprinted in Wax (1954).

Rosenblatt, F. (1958). The perceptron; A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**, 386–408.

Rosenblatt, F. (1962). *Principles of Neurodynamics.* Spartan Books, New York.

Rosenblueth, A., Wiener, N., and Bigelow, J. (1943). Behavior, purpose and teleology. *Philos. Sci.* **10**(1), 18–24.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning internal representations by error propagation. In Rumelhart and McClelland, 1986.

Rumelhart, D.E., and McClelland, J.L. (1986). *Parallel Distributing Processing.* MIT Press, Cambridge, MA.

Russell, B., and Whitehead, A.N. (1908). *Principia Mathematica*. Cambridge University Press, Oxford, England, 1962.

Sakurai, J.J. (1985). *Modern Quantum Mechanics*. Addison Wesley, New York.

Schoendorf, W.H., Marzetta, T.L., Perlovsky, L.I., Martinsen, E.A., Wallace, R.G., Rais, H., and Affens, D.W. (1994). *Application of a maximum likelihood algorithm to the detection of targets in SAR images*. *SPIE Conference on Optical Engineering in Aerospace Sensing*, Orlando, FL.

Schopenhauer, A. (1819). *The World as Will and Representation*. E.F. Payne, trans. Dover, Boston, MA.

Schrödinger, E. (1944). *What Is Life?* Cambridge University Press, Cambridge, England.

Searle, J. (1980). *Minds, Brains, and Programs, The Behavioral and Brain Sciences, 3*. Cambridge University Press, Cambridge, England.

Searle, J. (1992). *The Rediscovery of the Mind*. MIT Press, Cambridge, MA.

Searle, J. (1997). *The Mystery of Consciousness*. New York Review Press, New York, NY.

Searle, S.R. (1982). *Matrix Algebra Useful for Statistics*. John Wiley, New York.

Sebeok, T.A. (1972). *Perspectives in Zoosemiotics*. Mouton, The Hague.

Sebeok, T.A. (1977). *A Perfusion of Signs*. Indiana University Press, Bloomington, IN.

Segre, A.M. (1992). Applications of machine learning. *IEEE Expert* **7**(3), 31–34.

Shannon, C.E. (1948). The mathematical theory of communication. *Bell Syst. Tech. J.*

Shore, J.E. (1984). On a relation between maximum likelihood classification and minimum relative-entropy classification. *IEEE Trans. Info. Theory*, **IT-84**, 851–854.

Shore, J.E., and Johnson, R.W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Inform. Theory* **IT-26**(1), 26–37.

Simpson, P.K. (1990). *Artificial Neural Systems*. Pergamon Press, New York.

Singer, R.A., Sea, R.G., and Housewright, R.B. (1974). Derivation and evaluation of improved tracking filters for use in dense multitarget environments. *IEEE Trans. Inform. Theory* **IT-20**, 423–432.

Sittler, R.W. (1964). An Optimal Data Association Problem in Surveillance Theory. *IEEE Trans. Military Electronics* (**MIL-8**), 125–139.

Skinner, B.F. (1974). *About Behaviorism*. Alfred A. Knopf, New York.

Solovyov, V. (1985). Kant's philosophy. In Encyclopaedia, Brokgauz and Efron, St. Peterburg, Russia.

Spearman, C.E. (1904). General Intelligence Objectively Determined and Measured. *Am. J. Psych.* **15**, 201–293.

Specht, D.F. (1966). Generation of polynomial discriminant functions for pattern recognition. *Stanford Elec. Lab. Rep.* SU-SEL-66-029.

Specht, D.F. (1990). Probabilistic neural networks. *Neural Networks* **3**(1), 109–118.

Streit, R.L. (1995). Track initialization sensitivity in clutter. *Proceedings of the Signal and Data Processing of Small Targets, SPIE International Symposium*, San Diego, CA, **2561**, pp. 460–471.

Streit, R.L., and Luginbuhl, T.E. (1990). Maximum likelihood training of probabilistic neural networks. *NUSC Tech. Memorandum*, TM #911277. New London, CT.

Streit, R.L., and Luginbuhl, T.E. (1994). Maximum likelihood method for probabilistic multi-hypothesis tracking. *SPIE International Symposium on Signal and Data Processing of Small Targets*, *SPIE Proceedings* Orlando, FL, **2335**(24).

Stuart, C.I.J.M., Takahashi, Y., and Umezawa, H. (1978). Quantum neurodynamics. *J. Theor. Biol.* **71**, 605.

Sun, R., and Bookman, L.A. (1995). *Computational Architectures Integrating Neural and Symbolic Processing*. Kluwer Academic Publishers, Boston, MA.

Taborsky, E. (1998). The Architectures of Intelligent Systems. Conference on Intelligent Systems and Semiotics '98. Gaithersburg, MD.

Taborsky, E. (1999). Emotions As Forms of Consciousness. Joint Conference on Intelligent Systems ISIC/ISAS 99. Cambridge, MA.

Taylor, J.G. (1994a). Goals, drives and consciousness. *Neural Networks* **7**(6/7), 1181–1190.

Taylor, J.G. (1994b). Modelling what it is like to be. In *Toward a Scientific Basis for Consciousness*, Hameroff et al., eds. MIT Press, Cambridge, MA.

Thomas, D. (1995). Clutter characterization. *US/AS MoA Radar Research Meeting*, Adelaide, Australia.

Thurstone, L.L. (1947). *Multiple Factor Analysis*. University of Chicago Press, Chicago, IL.

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley, New York.

Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Welsey, Reading, MA.

Turing. A.M. (1937). On computable numbers with an application to the Entscheidunggsproblem. *Proc. London Math. Soc.* (*Ser. 2*) **42**, 230–265.

Vandrak, R.R., Smith, G., IIatfield, V.E., Tsunoda, R.T., Frank, V.R., and Perreault, P.D. (1977) *Chatanika Model for the HF Propagation Prediction*. RADC-TR-78-7, SRI, Menlo Park, CA.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* Springer-Verlag, New York.

Walker, A. M. (1964). Asymptotic properties of least squares estimates of parameters of the spectrum of a stationary non-determininstic time series. *J. Aust. Math. Soc.* **4**, 363–384.

Watanabe, S. (1985). *Pattern Recognition: Human and Mechanical*. John Wiley, New York.

Watson, J.B. (1913). Psychology as the behaviorist views it. *Psychol. Rev.*, **20**, 158–177.

Wax, N. (1954). *Selected Papers in Noise and Stochastic Processes*. Dover, New York.

Webb, J.C. (1980). *Mechanism, Mentalism, and Metamathematics*. Reidel, Dordrecht, The Netherlands.

Werbos, P. (1974). Beyond regression: new T. Ph.D. thesis, Harvard.

Wheeler, J.A. (1988). World as system self-synthesized by quantum networking. *IBM Journ. Res. & Dev.* **32**(1), 4–15.

Whorf, B.L. (1936). In *Language Thought and Reality: Selected Writings of Benjamin Lee Whorf*. J.B. Carroll (Ed.), MIT Press, 1964, Boston, MA.

Widrow, B. (1959). Adaptive sample-data systems—A statistical theory of adaptation. *WESCON Convention Record*, Part 4, 74–85.

Widrow, B. (1988). Efficiency of adaptive algorithms. *Twenty-Second Asilomar Conference on Signals, Systems & Computers*, Asilomar, CA.

Wiener, N. (1948). *Cybernetics*. John Wiley, New York.

Windelband, W. (1883). *The History of Philosophy*, J.H. Tufts, trans. Macmillan, New York.

Winston, P.H. (1984). *Artificial Intelligence*, 2nd ed. Addison-Wesley. Reading, MA.

Xenophanes. (VI BC). In *The Presocratic Philosophers*, G. Kirk and J.E. Raven, trans. Cambridge University Press, Cambridge, 1961.

Xu, L., and Yuille, A.L. (1995). Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Trans. Neural Networks*, **6**(1), 131–143.

Yarman-Vural, F., and Ataman, E. (1987). Noise, histogram and cluster validity for Gaussian-mixtured data. *Pattern Recog.* **20**(4), 385–401.

Yule, G.U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Trans. Royal Soc.* **A226**, 267–298.

Zadeh, L.A. (1962). From circuit theory to system theory. *Proc. Inst. Radio Eng.* **50**, 856–865. See also, Kandel, A. (1986). *Fuzzy Mathematical Techniques with Applications*. Addison-Wesley, Reading, MA.

Zadeh, L.A. (1965). Fuzzy sets. *Inform. Control* **8**, 338–352.

Zadeh, L.A. (1997). Information granulation and its centrality in human and machine intelligence. *Proceedings of the Conference on Intelligent Systems and Semiotics '97*, Gaithersburg, MD, pp. 26–30.

Zebker, H.A., van Zyl, J.J., and Held, D.N. (1987). Imaging radar polarimetry from wave synthesis. *J. Geoph. Res.*, **92**, 683–701.

Zeki, S. (1996). *A Vision of the Brain*. Blackwell, Oxford, England.

Zubarev, D.N. (1971). *Nonequilibrium Statistical Thermodynamics*. Nauka, Moscow.

Zue, V. (1996). *Spoken Language Systems*. Annual Research Summary, Laboratory for Computer Science, MIT, Boston, MA.

This page intentionally left blank

INDEX

A

A priori: and AI, 7; and Aristotle's Forms, 7–9; intelligent tracker, 9–11; internal models, 9; knowledge, xix, xxi, 3–4, 10; and language faculty, 7; and learning 4–6; and McCulloch, 4–5; measures of similarity, 10; and Minsky 5–6; neural structures, 4–6; and Plato's Eide, 4–5; similarity measures, 10
Abduction, 76; and analogy, 77
Abductive Reasoning, 76–77; and trees, 77
Adalines, 53
Adaptation. *See* Learning
Adaptive fuzzy association, membership, 9, 57
Adaptive fuzzy similarity. *See* Similarity measures
Adaptive resonance theory. *See* ART
Adaptive segmentation, 159
Adaptivity. *See* Apriority
Additive neural network, 58
Affect. *See* Emotions
Afferent signals, 8, 140
Afferent vs. efferent signals, 8, 112, 140
Agents, 8–9, 89, 94–95, 131, 161; CAS, 371; Kant-MFT, 361–363; MFT, 193, 207, 232, 314; and symbol, 376–377
Agrawala, 63
AI Debates: classical, 85–90; emerging, 91–94
AI. *See* Artificial intelligence
AIC, 193
Akaike information criterion. *See* AIC
Akhundov, M., xxi
Albus, J.S., xxi, 93, 101, 123
Aleshkovsky, U., xxi
Alternative choice states, 172; and information, 177–178; vs. intelligence and meaning, 178; number of, 178–179. *See also* Information
Ancient Greece, xx, 129, 135
Anderson, T.W., 46
Antiquity, xix-xxi, 4, 125, 128, 138
Antisthenes, 5, 126, 131
Apriority: vs. adaptivity of mind 4, 12–13, 129, 141–142, 144, 160, 194, 259, 371, 388, 419; vs. creation, 129
Aquinas,T., 7, 126, 131, 147; universals, 142
Archetypes, 141, 147n.21; timeless, 403. *See also* Jung, C.G.
Architecture, 4. *See also* Organization
Aristotle: 421; cause, end, 128; cause, formal, 128; combinatorial complexity of logic, 10, 53; controversy, 8, 13, 121n.2, 128; epistemology, 142, 145n.3; forms, xix-xx, 51, 126–127, 142; and fuzzy logic, 121; and Gödel, 389; learning, 4, 7, 127; logic, 51, 55; mathematics of mind, xx, 8, 119; ontology, 142, 145n.3; vs. Plato, 4, 7, 128; semiotics, 104; similarity, xx
ART, 111–113; distributed ART, 113; and emotions, 116; and MFT, 119, 258, 424
Artificial intelligence: and a priori knowledge, 68; combinatorial complexity, 69; general, 3; and learning, 68; and Plato, 69; rule-based, xix, 6
Artificial neural networks. *See* Neural networks
Assignment. *See* Association
Association, 3, 9, 37, 297; as assignment, 38; subsystem, 187, 244. *See also* CAT; MHT; Tracking
Attention, 314–315
Autoregression, 33–34; multidimensional, 291; non-linear, 294
Avicenna, 7, 130; on theology and philosophy, 147; universals, 142
Awareness, 369, 386, 403. *See also* Consciousness
AZ-similarity, 159, 181, 209, 293. *See also* Likelihood

B

Backpropagation, 64–66. *See also* Discriminating surfaces
Baggenstoss, P.M., 198
Bar-Shalom, Y., 42, 47, 123
Bartlett, M.S., 279
Bayes, Rev. T.: a posteriori probabilities, 20, 170, 184, 293, 322; decision theory, 1, 37; hypothesis testing, 3
Bayesian MFT. *See* MLANS
Bayesian networks, 77–78
Bayesian similarity, 166. *See also* MLANS
Beauty, 356, 363, 367, 399, 416; changing nature of, 411, 417; mathematics of, 367
Behaviorism, 133, 142; vs. cybernetics, 134; vs. mentalism, 134; and neural network crisis, 6

Being, 130
Bellman, R.E., 46, 52–53, 123
Berdyayev, 419, 421
Berg, R., xxi
Bergson, 356
Berwick, R.C., 123
Bhatnagar, 77, 122
Binford, T.O., 46, 52, 54, 74
Bishop, 123
Blackman, S.S., 47
Blum, A.L., 123
Böhme, 419
Bonnisone, P.P., 46, 54, 74
Bookman, L.A., 14
Boole, 121n.2, 143
Botha, R.P., 46, 123, 147, 421
Bottom-up processing, 74–75, 111
Boundary contour system, 114
Brain, 409–411; and Kant-MFT theory, 411–412
Brockett, R.W., 327
Brooks, R.A., 46, 52, 54, 74
Buddhism, xx; emptiness and instinct for knowledge, 395
Burdick, B.J., 260
Burg, J.P., 287
Burgeois, F., 47

C

Califano A., 46, 54, 74
Cantor, 385
Carpenter, G.A., 63, 111–112, 123, 421
CAS. *See* Complex adaptive systems
CAT, 297; Bayesian, 297–298; vs. CRB, 344–346; declaration, 304; examples, 308–311; Shannon-Einsteinian (SE) ANS, 304–306. *See also* CAT models; Einsteinian neural network; MLANS
CAT models: clutter, 301; Doppler, 302; linear, 299; link-track, 300–301; resolved object, 303; second-order, 300; spatio-temporal, 306–307
Chaldeans, 135
Chalmers, 395
Chang, W., xxi, 196
Chen, R.T., 46, 54, 74
Cherkassky, V. 79, 82, 123
Chomsky, N., 7, 69, 123, 126, 138, 142, 421; limits to scientific method, 139
Christianity, xx
Classification, 7, 212, 310; bounds, 86, 119, 192; degenerate geometry, 59–60; features, 20; and nominalism, 62; space, 23, 53; supervised, 228; unsupervised, 208. *See also* CRB; Learning
Clustering, 3, 56, 206–207, 222, 347
Clutter, 29, 243, 269, 281
Cognitive science, 146n.10, 378n.2
Collective consciousness, 396, 407
Collective unconscious, 130, 149n.21, 392, 419n.5
Competitive and cooperative dynamics, 44
Competitive neural network, 58
Complex adaptive systems, 107–108; vs. fuzziness, 109–110; and MFT, 368
Complexity: of CAT, 297; combinatorial complexity, 51, 121n.1, 121n.4; computational, 120; conundrum, 29, 52, 144; of MBV, 77; of MFT, 193; of nearest neighbor concept, 60; of problems, 80–81; of QMFT, 325; of rule systems, 7, 71–73. *See also* Simple and complex problems; Statistical Learning Theory
Computability, 120
Computation: computational model, 136, 146n.16; metaphor vs. model, 137; quantum, *See* Quantum computing
Computational concepts, 42; of modeling intellect 51.
Computational intelligence 121; vs. concepts of mind in philosophy, psychology, 51
Concepts, 4; conceptual signals, 118; and consciousness, 397; crisp concepts, 147n.22; and differentiation, 398; emotional concepts, 398; fuzzy concepts, 147n.22; general 5
Concepts of intellect: mathematical, 3, 43, 71, 85–86; of mind, 143–145; philosophical, 3, 11, 43, 71, 129, 147n.15
Concurrent association and tracking. *See* CAT
Conditioned stimuli, 116
Connectivism, 92
Consciousness: of bodhisattvas, 395; collective, individual, and individualistic, 400–401; conceptions and misconceptions, 392–393; and Ego, 394, 404; emergence in history, 401–402; and emotions, 397; and differentiation, 393–395, 398, 404; and fuzziness, 396; hard and easy questions, 395; intentionality, 406; levels, 393; modalities, 393, 399, 404; mysteries, 418; neural structures, 409–411; purpose, 400; and resonance, 115; and Self, 394; time, 403; and unconscious, 396; and Understanding, 396. *See also* Jaynes, J.; Maimonides, M.; Searle J.; Unconscious
Contradiction law, 102
Conundrum of combinatorial complexity, 29, 52
Convergence, 188, 190, 238, 254; global, 172, 188, 256–257; local, 254. *See also* Local maxima
CR Theorem, 331–332
Cramer, H., 16, 149, 198, 353
Cramer-Rao bound. *See* CRB
Cramer-Rao theorem *See* CR theorem
CRB, 329; for CAT 344–346; classical, 330; for intellect and evolution, 348–349; for General MLANS, 333–334, 339–342; and learning, 329–330. *See also* Fundamental bounds
Creativity, 4; and beauty, 417; and consciousness, 417; and differentiation, 416–417; and free will, 415–416; mysteries of, 418; and unconscious, 417
Crevier, 123
Cross, 63
Curse of dimensionality, 53, 81, 251
Cyberaesthetics, 364, 417
Cynic school of philosophy, 5, 131

D

Dalai Lama, 395
Data mining, 81, 99, 295–296
Daugman, J.G., 137, 147
Daum, F.E., 353, 354
Debates, mathematical, 4–5, 42, 85–86; apriority vs. adaptivity, 85; cognition vs. perception, 89; consciousness, 94; emotions, 93; hierarchy vs. heterarchy, 93; holistic, 91; hybrid systems, 86; logic vs. neural 85, 90; molecule as a sign, 92; non-algorithmic, 90; non-intelligent agents, 89; parallel organization, 90; representation 87–88; sequential computation, 90; signs vs. symbols, 92; thinking, 86–87; thinking vs. acting, 91; understanding, 87; what vs. how, 91. *See also* Consciousness
Debates, philosophical, 4–5, 42
Debates, theological, 5, 42
Decision theory, 19. *See also* Bayes, Rev. T.
Declaration, 302, 307–308
Deduction, 76
Degenerate Geometries, 58, 62, 69
Deity. *See* God
Dempster, A.P., 198
Descartes, 130, 132, 135, 138
Designatum, 11–12
Differentiated phenomenology of consciousness, 45, 388, 391, 394, 399, 404, 409
Discriminating surfaces, 62–63; combinatorial complexity, 68; and neural networks, 63–64; and nominalism, 67; training requirements, 67–68
Dmitriev, V.A., xxi, 421
Dreyfus, on representation, 88
Drive. *See* Instinct
Duda, R.O., 46, 53, 55, 63
Durbin, J., 287
Dyer, C.R., 46, 54, 74
Dynamical symbol. *See* Symbol

E

Eco, U. 105
Efferent signals, 131. *See also* Afferent vs. efferent signals
Efficiency of learning and estimation, 260, 278, 330
Ego, 141, 388, 391, 398
Eide, xix, 4, 70. *See also* Ideas
Einstein, A., 125, 137, 147, 172
Einsteinian likelihood, 172–173. *See also* Shannon-Einsteinian similarity
Einsteinian mixture model, 174
Einsteinian neural network, 175; agents, 184; dynamics, 183–4; equations, 272; and photon ensemble, 176; spectral models, 270; submodels, *See* agents; time-frequency models, 272–273
Einsteinian neural network examples: 1-D spectra, 274–277; OTH spectra, 282; 2-D time-frequency, 278–279. *See also* CAT
EM Algorithm, 174, 190, 256
Embedding fields, 111
Emergence of consciousness, 400–401, 414–415
Emotional intellect, 356, 369–370
Emotional machine, 366–367, 370; and beauty, 367
Emotions, 3, 130, 356, 364; vs. affect, 398; differentiated and nondifferentiated, 398; emotional signals, 118; higher, 368–369; and learning 116; neural structures, 412; and rationality, 412
Empirical risk minimization, 82–83
Empiricism, 5, 131–132, 143
Engineering applications, 4, 42, 107, 152, 367; and philosophy, 8
ENN. *See* Einsteinian neural network
Ensemble, 173–175, 180, 186, 283
Entropy, 176–177; vs. information, 197n.7. *See also* Information
Epistemology, 142, 145n.3
Ergodicity, 186
ERM. *See* Empirical risk minimization
Estimation system, 176, 185, 192
Evolution, bounds on, 326, 348; of brain, 87; of consciousness, 45, 94, 100; of intelligence, 92. *See also* Evolutionary computation; Learning
Evolutionary computation, 106–107, 123
Excluded third law, 7, 359. *See also* Contradiction law
Expectation-maximization algorithm, 192
Expert systems. *See* Rule-based systems

F

Factor analysis, 134, 146n.12–14
FCS. *See* Features contour system
Features, 6, 25, 101, 155, 208, 213, 260, 276, 344
Features contour system, 114
Feedforward neural networks, 63–64; learning disability, 67; mathematical formulation, 64
Feelings. *See* Emotions
Feynman, R.P., 327
Financial market forecast, xx, 34–35, 41, 267, 292
Fogel, 123
Formalism, 382
Forms as potentialities, 7, 128, 142, 161, 165, 171, 193, 258
Forms of Aristotle. *See* Aristotle, forms
Fortmann, T.E., 47
Forward and inverse problems, 22, 186, 263
Fossum, 63
Fourier transform, 266
Frames, 71, 97
Franchi, P.R., 287
Frank-Kamenetsky, M., 378
Franklin, 123, 423n.1; computational complexity, 70; thinking, 88
Free will, 4, 45, 386, 389, 392, 397, 408, 418, 421; and behaviorism, 133; and causality, 361; and consciousness, 399, 401, 403; and creativity, 411, 415; and human-in-itself, 418; Maimonides, 399; the mystery of, 418; Penrose, 381; and unconscious, 394
Freeman, W.J., 48, 112, 359, 378, 421
Frege, 386
Freud, S., 86, 93, 139, 356; cathexis vs. neuronal

excitation, 140; realism and nominalism, 140; romanticism of, 420; structures of psyche, 140; unconscious, 140
Fukunaga, K., 53, 260
Fukushima, 58, 123
Fundamental bounds, xx; on learning, 206, 222, 328, 348. *See also* CRB
Fundamental computational concepts, xx, 43, 51, 164
Fundamental limitations. *See* Fundamental bounds
Fundamental mathematical concepts, xx, 43, 51, 164
Fundamental philosophical concepts, xx, 129–130
Fuzzy concept, 10, 12, 55, 90, 148n.23, 165; and differentiation, 394; and MFT, 193, 361, 364; and symbol, 374
Fuzzy logic, xx, 8, 10, 13; control of fuzziness, 197n.5; membership, 57, 157, 160, 170, 174, 187, 190, 201, 310

G

Galen, 412
Gallileo, 137
Gardner, H., 123
Garvin, L.C., 196, 261, 328
Gated dipole, 117
Gaussian mixture model, 174–175, 184, 238, 241–243, 284, 293–294
General fuzzy non-linear autoregression, 294
General fuzzy non-linear regression, 292–293
Generalization, xix, 69, 80, 85, 194
Genetic algorithms, 106–107, 371–372; and MFT, 372–374
Genetic operators, 108
GFRANS, 292
Gibbs Quantum Modeling Field System, 325
Girosi, F., 47, 52, 123
Global convergence. *See* Convergence; Local maxima
Gnostics, 44, 149n.15, 412
Goals, 4, 71, 101, 119, 231, 314, 357–358, 373; and beauty, 421; and consciousness, 397, 400; and differentiation, 417; goal-directed functioning, 4, 409; and intentionality, 402–403
God, 5, 128, 419; adaptivity, 145n.4; anthropomorphic, 129; design principle, 365; emanation, 129; and free will 399; as realization of idea, 415; as spirit, 129; as thought, 129; transcendence vs. immanence, 130, 145.n4
Gödel, K., 93, 421; theorems, 55, 387; theory, 385
Goldfarb, L. 47
Goodman, L.E., 149
Gradient learning, 53, 63
Graham, M.L., 353
Greineder, S., 198
Grimson, W.E.L., 46, 52, 54–55, 74
Grossberg, S., 7, 58, 91, 122–123, 126, 140, 400, 409, 421; additive neural network, 258; Aristotelian mathematics, 8; ART 90, 111–112; attention, 315; consciousness, 115; emotions, 117; gated dipole, 117; and Kant, 114; and Maimonides, 130; motor control, 115
Grossberg's method, 111. *See also* Kant-Grossberg method
Grouping. *See* Clustering
Growth of knowledge, 147n.22
Gudwin, R., 93

H

Hall, D.L., 122
Hameroff, S.R., 124, 424
Hamiltonian Quantum Modeling Field System, 326
Hart, P.E., 46, 53
Haykin, S., 123
Hebb, D., 5
Hebbian learning, 258–261
Hegel, 147n.15, 356
Helmholtz, 136, 254
Heterarchical architecture, 7, 44, 93, 112, 161–162, 358, 363–364
Hierarchical ENN+MLANS Architecture, 280
Hierarchical organization of intelligent systems, 9, 101–102; elements of intelligence, 103; Kant-MFT, 361; NIST model, 102–103
Higher emotions, 119, 360, 366–367, 373, 376n.2, 396. *See also* Instinct
Hilbert, 120, 385
History: of mathematical concepts of intellect, 3, 43, 71, 85–86; of philosophical concepts of intellect, 3, 11, 43, 71, 129, 147n.15. *See also* Concepts of intellect
HMD, 198
HME, 198
Ho, 63, 122
Holland, J.H., 108, 378, 421
Hopf, L.,122, 172
Hopfield, J., 58, 122
Horgan, 88
Housewright, R.B., 49
Hume, 134, 142
Huttenlocher, D.P., 52, 55
Hypothesis choice. *See* Bayes, Rev. T.; Decision theory

I

Ideas, 4; of Plato, xix, 7, 69–70, 74, 102, 129, 193, 359
Illusions, 114
Image processing: classification, 59, 240; recognition, 25, 44, 74, 152, 172, 276, 304; understanding, 349. *See also* MBV
Immanence vs. transcendence, 44, 130, 140, 146n.4, 419
Individuation, 397, 404
Induction, 76
Information, 172; and alternative choice states, 177; and Einsteinian likelihood, 177, 183; vs. entropy, 197n.7; likelihood, 186; mutual information, 180
Information fusion, 312–313
InnoVerity.com, xix
Instinct, 112, 118, 194, 368–369, 395
Intellectual emotions, 368–369
Intelligence, 4, 7, 10–12, 25, 43, 70, 89–93, 110. *See also* Agents; Artificial intelligence; Debates, mathematical; Mathematical concepts of intellect
Intelligent agent. *See* Agents

Intelligent systems: 4, 44, 77, 369, 379; architecture, organization, 3, 105; and beauty, 421; and CRB, 349; and Gödel, 385; and Kant, 357; rule-based, xix; and symbols, 92. *See also* Artificial intelligence; Modeling Field Theory; Neural networks
Intelligent tracker, 9–12
Intentionality, 406–407; and beauty, 368; and causes of Forms, 406; of consciousness, 406–407; and purposiveness, 367
Internal models. *See* Models, internal
Interpretant, 11–12, 104, 373, 375, 376n.5. *See also* Semiotics; Symbol
Interpreter, 11–12, 104. *See also* Semiotics; Symbol
Intuition, 4; mysteries of, 418; physical, 4, 418
Inverse problems. *See* Forward and inverse problems
Ionosphere: clutter, 282; e-m wave propagation, 281. *See also* Radar
Islam, xx, 130

J

Jacobs, R.A., 198
Jaffer, A.J., 47
Jain, A.K., 46, 52, 54, 74, 90
James, W., 403
Jaspers, 356
Jaynes, E.T., 184; 198
Jaynes, J., 147, 149, 424; differentiated consciousness, 404; emergence of consciousness, 401; individual consciousness, 402–404; misconceptions about consciousness, 392
JDL model, 97–98; hierarchy, 99
Joint Directors of Labs model. *See* JDL model
Jordan, M.I., 198
Judaism, xx
Judgment, 359–360; and differentiation, 399; and similarity, 360
Jung, C.G., 93, 126, 142; archetypes, 141, 147n.21; individuation, 404; projection, 135, 139; unconscious, 141

K

Kalam, 129
Kalman filter, 36
Kanal, L.N., 46, 77, 79, 122, 123
Kant, I.: and Aristotle, 7; beauty, 123; content of a priori knowledge, 138; pure spirit, 138; and semiotics, 376–377; on space perception, 87; synthetic judgments, 145n.8; theory of mind, xix-xx, 93, 123, 356; thing-in-itself, 139; universals, 142. *See also* Judgment; Pure reason; Reason; Understanding
Kant-Grossberg method, 114, 421
Kant-MFT theory, 361, 376, 385
Kay, 287
Keshavan, H.R., 46, 74
Klopf, A.H., 260
Knowledge instinct, xix, xxi, 94, 97, 116, 119, 151, 194–195, 356, 367–368, 397, 424n.4
Kohonen, T., 58
Koster, J., 46
Kryukov, V.I., 258
Kumar, V.P., 198

L

Lamdan, Y., 46, 54, 74
Landau, L.D., 198
Lang, 63
Language faculty, 7, 69, combinatorial complexity, 70
Lassale, J.-C., 47
Learning from experience, 5; vs. programming of structures 192–193, 231
Learning, xix, xxi, 3–7, 9, 52, 53, 101, 140; Aristotle, 129, 169; backpropagation, 66; vs. beauty, 365, 399, 421; in CAS vs. MFT, 371; complexity of, 29, 43; and consciousness, 401; vs. convergence, 254; and creativity, 422; disability, 67; efficiency of, 296; and emotions, 116; evolutionary, 106; and fuzziness, 51–52, 169, 172, 296, 422; vs. goal, 236; and Gödel theorem, 386; gradient learning 64; vs. hierarchy, 94; and information, 177, 179; and intentionality, 361, 407; limitations of, 326–327, 349; in linguistics, 88, 127; MFT, 187; MHT, 76; in MLANS, 169, 210; of models, 367; partially supervised, 228; rule-based, 69; and nominalism, 142; in Soar, 72; structures, 191–193; supervised 25; vs. symbol, 358, 375; with teacher, 233; unsupervised, 215, *See also* Clustering. *See also* Abduction; Apriority; ART; Artificial intelligence; Induction; Modeling Field Theory; Nearest neighbor; Neural networks; Statistical Learning Theory
Learning instinct, xix, xxi, 94, 97, 116, 119, 151, 194–195, 356, 367–368, 397, 424n.4. *See also* Knowledge instinct
Leibler, R.A., 184
Leibnitz, 132
Levine, D. S., 123, 412
Levinson, 287
Lévy-Bruhl, 403
Lifshitz, E.M., 198
Likelihood, 21–22, 27, 153, 167; for association, 38; ratio test, 21; and regression, 32; and segmentation, 169; and similarity, 162, 168. *See also* Einsteinian likelihood; Maximum likelihood; MLANS; Shannon-Einsteinian similarity
Limits on learning. *See* CRB; Learning
Linguistics, 7, 69, 127, 371; combinatorial complexity, 70. *See also* Chomsky, N.
Local maxima, 164, 189. *See also* Convergence
Logic, 121n.2; and Gödel theorem, 388–389; and mind, 385
Locke, 134, 142
Longstaff, 63
Long-term memory, 111, 258
Lorenz, K., 395
Lozano-Perez, T., 52, 54, 74
LTM. *See* Long-term memory
Luginbuhl, T.E., 198

M

Maimonides, M., 7, 147; and Aristotle, 130; on collective and individual consciousness, 402; finite angels, 130; monotheism and adaptivity, 130; universals, 142
Manolakos, E.S., 198
Martinsen, E.A., 260
Marty, K., 47
Materialism vs. idealism, 42
Mathematical concepts of intellect, 1, 6–7, 43, 51, 53, 98, 126, 129, 135, 364, 388; classical 3, 51; theory of emotional intellect, 356–358, 368
Mathematics of beauty, 356, 365, 407
Mathematics vs. philosophy. *See* Concepts of intellect
Mathematics vs. physics, 135–136
Maximum Entropy Adaptive Neural System. *See* Einsteinian neural network
Maximum entropy. *See* Maximum information
Maximum information, 152, 185–186, 270, 277. *See also* Einsteinian neural network Maximum likelihood, 28, 41, 152, 166, 191, 235, 252, 270. *See also* MLANS
Maximum Likelihood Adaptive Neural System. *See* MLANS
Maxwell, 136
May, R., 46
MBV. *See* Model-based vision
McClelland, 90
McCulloch, W.S., 4, 142; a priori structures, 53; additive neural network, 122; realism, 5; revolution 6
McManus, M., 260
Meaning, 10, 104–105, 130, 358, 361, 376, 378n.5, 387, 406; vs. information, 178
MEANS. *See* Einsteinian neural network
Mental, 5, 90, 117, 132, 179, 387, 406, 421, 423n.1
Mentalism, 146n.10
Metric, 57, 206; class-dependent, 259; Eucledian, 59. *See also* Similarity measures
Metropolis N., 260
Meystel, A., 93, 101, 103, 123, 378, 390, 421
MFT. *See* Modeling Field Theory
MHT, 75–76, 113, 186, 353
MI. *See* Information
Michalski, R.S., 54, 74
Middle Ages, 4, 53, 126, 129, 130, 132, 147n.15
Mind vs. brain, 138
Minsky, M.L., 44n.1, 142, 421; agents, 94–95; AI debates, 85; delineating myths and limits, 96; learning in rule systems, 69; neural networks 5; realism, 54, 126; rule systems xix, 55; society of mind, 94–95; thinking 88
ML. *See* Maximum likelihood
MLANS, 168; agents, 171; architecture, 208–209; association subsystem, 187, 211; convergence, 213, 238; examples, *See* MLANS examples; fuzzy or probabilistic associations, 170; learning efficiency, 260; learning equations 169; likelihood, 210; modeling subsystem, 187, 211, 244; vs. other neural networks, 257–259; overview, 206; phase transitions, 258; structure learning, 231, 233–236. *See also* Modeling Field Theory
MLANS examples, 213, 219, 222; comparison to nearest neighbor, 222–224; complexity, 219–202; convergence sensitivities, 217–218; vs. CRB 342–343; interactive learning, 229; partial supervision, 228; SAR, 245–250; scene segmentation, 238; structure learning, 232; supervised-unsupervised combined learning, 225–227; unsupervised learning, 216; teacher, 227
Model-Based Neural Networks, xxi, 192. *See also* Einsteinian neural network; MLANS; Modeling Field Theory
Model-based recognition, 27–28; combinatorial complexity, 26, 29, 53; vs. Gaussian assumption, 45n.8, vs. independence assumption, 45n.9
Model-based vision, 74; adaptivity, 75; and MHT, 75. *See also* Model-based recognition
Modeling Field Theory, xix, 119; agents, 163–164; agents' fuzziness, 171; algorithm, 187; architecture, 187; and Aristotelian mathematics, 119; and ART, 119, 258, 424; Bayesian MFT, 165–166; and beauty, 367–369, 420; and CAS, 374–375; conceptual overview, 153, 161, 194–195; of consciousness; continuous, 163; convergence, 188, 190; dynamic equations, 162; EM algorithm, 192; and genetic algorithms, 373–374; Judgment, 359–360; learning instinct, 119, 194; local maxima, 164; organization heterarchical, 163–164; Reason, 361; and semiotics, 119, 375–376; and symbol, 376; Shannon-Einsteinian MFT, 172–174; Understanding, 357–359. *See also* CAT; Einsteinian neural network; Emotions; MLANS; Models, internal; Pure reason; Reason; Similarity measures; Thinking; Understanding
Models, internal: adaptive, xix; adaptive fuzzy 8, 10; antinomial nature, 420; approximate, 196; AR, MA, ARMA, 264–265; CAT models, 299–301; clutter, 241–242; dynamic, 33–36; motor control, 115; origins of, 372–373; outliers, 242–243; as representations, 154; Rician, 243–244; SAR and radar image, 238; sensory domain, 154; spectral, 267, 270, statistical; three aspects of, 378n.5; three levels, 154; Wishart, 241–242
Mohan, C.K., 46, 74
Monti, J., 196
Moore, K., 63
Moravscik, E.A., 123

Morris C., 104–105, 124
Motor coordination, 115
Mulier, F., 79, 82
Multilayer neural networks, 63–66
Multiple Hypothesis Testing. *See* MHT
Multiple Hypothesis Tracking, 39; combinatorial complexity, 40. *See also* MHT
Munkres, J., 47
Muratet, M., 313
Mysteries of physics, 417–418

N

Nahi, N.E., 47
NASA, 241–242, 260
Nearest neighbor concept, 53, 58; mathematical formulation, 59; and neural networks, 58
Negahdaripour, S., 46, 52, 54–55, 90
Neocognitron, 58
Neoplatonics, 129
Neumann von, J., 90, 327
Neural fields: nonlocal and nonstationary, 136; theories, 111. *See also* ART; Modeling Field Theory
Neural networks, 4, 51, 53, 101, 110, 161; early 6; crisis, 142; limitation on learning, 223. *See also* Adalines; ART; Backpropagation; Einsteinian neural network; Embedding fields; Feedforward neural networks; Gated dipole; GFRANS; HME; MLANS; Modeling Field Theory; Model-Based Neural Networks; Perceptrons; PNN; POEM; RCE neural network; SPNN
Neural vs. symbolic, 44
Neural weights, 64, 162, 189, 258
Neuron, formal, 5, 143
Nevatia, R., 52
Newell, A., 86; AI history, 123; on representation, 88
Newton, I., 132, 136–7, 147, 172, 369, 384, 413, 419
Nichols Research, xix
Nietzsche, F., 141, 147, 356; on free will, 421n.1
Nilsson, N.J., 46, 53, 55, 63
NNC. *See* Nearest neighbor concept
Nominalism, 5, 44, 62, 64, 74, 77, 85, 126, 131, 142, 388
Non-linear General Fuzzy Regression ANS, 292–293
Non-linear Regression, 292–293
Nonparametric methods, techniques, 23, 53, 74, 260, 267; and combinatorial explosion, 74
NRC, xix

O

Objectification, 145n.7
Occam, W., 4–5, 147; razor 111; universals, 142
Ontology, 142
Organization: of brain, 4; of an intelligent system, 3, 39, 91, 151, 161, 356, 404, 420. *See also* Brain; Heterarchical architecture; Hierarchical organization of intelligent systems; Intelligent systems
OTH, 279–281, 307
Over-the-Horizon Radar. *See* OTH

P

Papert, 6; AI debates, 85
Papez, J.W., 412
Parameter estimation system, 9, 176, 208
Parametric methods: in linguistics, 70; structures, 371–373, 419; techniques, 8, 23, 53–54, 76, 91, 115, 144, 155, 174, 189, 206, 349
Parametric vs. non-parametric estimation, 73; computational complexity, 74
Parzen method, 45n.6
Pattern recognition, statistical, 7, 22–23; a priori information, 24–25; combinatorial complexity 7, 53. *See also* MLANS
Pavlov, 133
PDA, 42
Peirce, C.S., 105, 124, 375
Penrose, R., 387, 421; on Gödelizing Turing machine, 388; on Ideas, 418; new physical phenomena, 384, 420; noncomputability of understanding, 383–384, 389; representation, 88, 123
Perceptrons, 6, 53
Perlovsky, L.I., 47, 52–53, 55, 64, 124, 198, 260–270, 287, 327, 424
Personality, 405. *See also* Psychic functions
Philosophy vs. mathematics. *See* Concepts of intellect; History; Mathematical concepts of intellect
Photon ensemble, 173–175. *See also* Ensemble
Physical intuition, 90, 413–414
Physics: Aristotelian, 128; of beauty, 45; of brain, 138; and computability, 120; of consciousness, 414; and Gödel theorem, 388; of material substance, 91; vs. mathematics, 136; of mind, 12, 48, 110–111, 113, 126, 137, 139, 368, 421; new, unknown, 90, 369, 422; of spiritual substance, 136, 138, 412; of symbol, 414. *See also* Kant-MFT theory
Pitts, W., 5, 53, 122
Plato, xix, 4, 126, 128, 131, 421; ontology vs. epistemology, 142, 145n.3
Plato-Minsky approach, 6, 91, 127
Plotinus, 129
Plum, C.P., 198
PNN, 198
POEM, 198
Poggio, T., 58, 63, 123
Pragmatics, 104, 375, 520. *See also* Semiotics
Prediction, 3, 14, 21, 25, 28, 31, 40, 55, 81, 102, 136, 289–290, 330
Pribram, K., 124, 376, 424
Proakis, J., 287
Probabilistic Data Association. *See* PDA
Probabilistic neural network, 58
Probability: basic notions 13–19; Bayesian hypothesis choice, 20–21
Production Systems, 71–73, 89
Psychic functions, 393, 395, 397, 406; and differentiation, 412, 415
Psychology of philosophy, 139
Pure reason, 357–358

Pure Spirit, 356
Putnam, H., 387–388
Purposiveness, 360, 366, 384, 421; and beauty, 375
Pythagoras, 135

Q

QMFT. *See* Quantum Modeling Field Theory
Qualia, 388, 393, 407
Quantum computing, 120, 321–322
Quantum measurement, 45, 173, 325, 382, 391, 418
Quantum Modeling Field Theory, 322–323; Gibbs QMF, 324–325; Hamiltonian QMF, 326–327
Quantum neurodynamics, 118; QBD, 118–119
Quidditive vs. existential aspects, 131

R

Radar, 30, 38, 172; image classification 204, 207, 245–247; spectra, 277–280
RCE neural network, 58
Realism, xix, 4, 70, 126, 131, 142; vs. nominalism, 126, 132, 388
Reason, 361. *See also* Kant, I.; Modeling Field Theory
Reduced coulomb energy. *See* RCE neural network
Regression: autoregression, 33–34, 291; data mining example 295–296; as expectation 32; GFRANS, 292; linear, 30–31, multidimensional, 290; nonlinear, 292; nonlinear autoregression, 294
Regularization neural network, 58
Reid, 47
Reilly, 58, 123
Resonance, in ART, 111; and consciousness, 115
Revenue prediction, 295–296
Riasanovsky, N., 419
Rice, S.O., 260
Rivest, R.L., 123
Romanticism, 419–420
Rosenblatt, F., 5
Rosenblueth, A., 134
ROTHR. *See* OTH
Rubin, D.B., 198
Rule-based systems, xix, 6, complexity, 7; and learning, 6–7. *See also* Artificial intelligence
Rumelhart, D.E., 90, 123
Russel, B., 55, 86, 121n.2, 143, 385

S

Sakurai, J.J., 198, 327
SAR, 238–239; data, 239–240; models, *See* Models, internal
Schema, 109, 370, 421; and fuzziness, 110, 371
Schoendorf, W.H., 196, 260
Scholasticism, 145n.6
Schopenhauer A., 357, 395
Schrödinger, E., 118, 325, 327, 383
Scientific method, 5, 89, 133, 357, 391, 418, 420
Sea, R.G., 47
Searle, J., 390; Chinese room, 87; consciousness, 392, 408; consciousness vs. MFT, 404; modalities of consciousness, 393; thinking, 123
Searle, S.R., 84, 200, 253
Sebeok, T., 105
Segre, A.M., 46, 54–55, 74
Sejnowski, T.J., 258
Semantics, 104, 375, 520. *See also* Semiotics
Semiosis, 375–376
Semiotics, 11–12; basics, 104–105; and Judgment, 376; and Reason, 376; sign vs. symbol, 106, 375–376; symbolic AI, 106; terminology, 44n.2; and thinking, 376; and Understanding, 376
Sensor fusion, 97; types of, 98. *See also* JDL model
Sensorimotor control, 115
Shannon, C.E., 87, 153, 165, 172, 183, 252. *See also* Shannon-Einsteinian similarity
Shannon-Einsteinian MFT. *See* Modeling Field Theory
Shannon-Einsteinian similarity, 173. *See also* Similarity measures
Shore, J.E., 186
Short-term memory, 111, 258
Sign, 11–12, 92, 375, 415. *See also* Semiotics
Similarity measures, 12, Adaptive fuzzy, 159; Aristotelian, 157; A-similarity (see Aristotelian); AZ-similarity (see Aristotelian Fuzzy); Bayesian A-similarity, 165–166; Bayesian AZ-similarity, 168–169; complexity, 158, 160; conditional, 156–157; fuzzy, 158; and Judgment, 359; Shannon-Einsteinian, 181–183. *See also* Entropy; Information
Simon, H.A., 86; logic, 88
Simple and complex problems, 59, 80–81, 84. *See also* Statistical Learning Theory
Simpson, P.K., 47, 58, 123
Singer, R.A., 47
Skinner, B.F., 142, 145n.9
SLT. *See* Statistical Learning Theory
Smith, A.F.M., 373
Soar, 70; learning, 71; preferences, 71, 94; productions, 72; rule systems and combinatorial complexity, 71
Society of agents, 95–96
Socrates, 126, 132, 419
Spatio-Temporal Patterns, 35, 144, 187, 259, 305
Spearman, C.E., 134
Specht, D.F., 58, 63
Spectrum estimation, 266–267. *See also* Einsteinian neural network
Spectrum, 265,
Speech recognition, 58, 156, 268, 276, 287. *See also* Transients signals
Spiritual processes, 5; substance, 4, 136
SPNN, 198
Statistical Learning Theory, 79–80, 192; empirical risk minimization, 80–81; VC dimension, 81–82; SVM, 82–83; generalization error, 84
STM. *See* Short-term memory
Stockbridge data, 246–247
Streit, R.L., 198, 353, 354
Strong AI, 146n.18
Stuart, C.I.J.M., 124
Sufficient statistics, 45n.7
Sun, R., 144

Support Vector Machines, 85. *See also* Statistical Learning Theory
Symbol, 7, 12, 25, 45, 93, 391; and AI, 69, 72, 92; and consciousness, 414; dynamic, 94, 104–106, 358, 375; process, 105, 375, 384, 414–415; and semiotics 11–12, 104–106, 122.n13; terminology, 42; and thinking, 396; and unconscious 414–416
Symbolic AI: as misnomer, 46n.1, 69, 92, 104–105, 375
Syntactics, 104, 375, 520. *See also* Semiotics
Synthetic aperture radar. *See* SAR
Synthetic judgments a priori, 358; and hierarchical models, 359
Systems of logical rules. *See* Artificial intelligence

T

Taborsky, E., 92, 424
Target tracking. *See* CAT
Taylor, J.G., 410, 424
Terzopoulos, D., 196
Theology, 129, 395; Buddhism, 395; creation, 129; monotheism, 129–130; pagan, 129; schism with philosophy, 129. *See also* Aquinas, T.; Avicenna; Dalai Lama; Kalam; Maimonides, M.; Occam, W.
Thing-in-itself, 138–139
Thinking, 3, 69, 102, 357, 370, 404; and semiotics, 10, 376
Thomas, D., 287
Three-component mathematical structure of the modeling field theory. *See* Kant-MFT theory
Three-component structure of mind. *See* Kant, I.
Thurstone, L.L., 134
Tichovolsky, E.J., 285, 287
Tienson, representation, 88
Time: and consciousness, 403; perception, 403
Time travel, 383
Time-Frequency 268, 272–273, 278–279
Titterington, D.M., 260
Top-down and bottom-up processing 75–76, 111–112, 363.
Tracking, 35–36; models, 36. *See also* CAT
Transcendence. *See* Immanence vs. transcendence
Transient signals, 116, 269
Trees, 80
Tse, E., 47
Turing, A.M., 77, 87, 138; test, 90; theory, 120, 387
Tye, D.M., 262

U

Umezawa, H., 118
Uncertainty, 154, 186
Unconscious, 45, 94, 104, 133, 395, 403, 420; and adaptation, 397; archetypes, 373, 395, 400, 424; collective 130, 139, 356, 367, 395, 398; and consciousness, 386, 389, 392–3, 408–409; and creativity, 417, 419; and differentiation, 394; and free will, 392; and fuzziness, 373; human-in-itself, 363; individual, 367; and intuition, 418; mathematical theory, 367; and projection, 137; and resonance, 115; and symbol, 373
Understanding, 357–368
Unity of apperception, 97, 405
Universal concepts, 5, 142; archetypes, 141; forms, 130. *See also* Models, internal
Universals. *See* Universal concepts

V

Vapnik, V., 46, 79, 81–82, 84
VC-dimension: 81–85; definition 81–82, examples 83

W

Walker, A. M., 287
Wallace, R.G., 373
Watanabe, S., 46, 53
Watson, J.B., 149
Webb, J.C. 381, 387
Werbos, P., 68, 123, 124
Weyl, 385
Wheeler, A., 105
Whitehead, A.N., 86, 390
Whorf, 403
Widrow, B., 5, 287
Wiener, N., 53, 134, 286; realism and nominalism, 134; theology, 134
Wigner, 383
Will, 361, 376. *See also* Reason
Windelband, W., 147
Winston, P.H., 46, 47, 52, 54–55
Wirth, J.R., 123
Wishart model. *See* Models, internal
Wolfson, H.J., 46, 54

X

Xenophanes, 129

Y

Yuille, A., 327
Yule, G. U., 278, 286–287

Z

Zadeh, L.A.: fuzzy logic, xx, 8, 13, 121n.2, 143–144, 147, 378, 421; granularity 103; limitations of, 56. *See also* Similarity measures
Zeki, S., 408
Zubarev, D.N., 327
Zue, V., 287